

Testing Forest-Isomorphism in the Adjacency List Model

Mitsuru Kusumoto^{1*} and Yuichi Yoshida^{1,2**}

¹ Preferred Infrastructure, Inc. mkusumoto@preferred.jp

² National Institute of Informatics. yyoshida@nii.ac.jp.

Abstract. We consider the problem of testing if two input forests are isomorphic or are far from being so. An algorithm is called an ε -tester for forest-isomorphism if given an oracle access to two forests G and H in the adjacency list model, with high probability, accepts if G and H are isomorphic and rejects if we must modify at least εn edges to make G isomorphic to H . We show an ε -tester for forest-isomorphism with a query complexity $\text{polylog}(n)$ and a lower bound of $\Omega(\sqrt{\log n})$. Further, with the aid of the tester, we show that every graph property is testable in the adjacency list model with $\text{polylog}(n)$ queries if the input graph is a forest.

1 Introduction

In *property testing*, we want to design an efficient algorithm that distinguishes the case in which the input object satisfies some property or is “far” from satisfying it [11]. In particular, an object is called ε -far from a property P if we have to modify an ε -fraction of the input to make it satisfy P . A (randomized) algorithm is called an ε -tester for a property P if it accepts objects satisfying P and rejects objects that are ε -far from P with high probability (say $2/3$).

Graph property testing is one of the major topics in property testing, and many properties are known to be testable in sublinear time or even in constant time (in the input size). See [5] for surveys. In order to design sublinear-time testers, we have to define how to access the input graph, as just reading the entire graph requires linear time. The model used here is the *adjacency list model* [9]. In this model, the input graph $G = (V, E)$ is represented by an adjacency list and we are given an oracle access \mathcal{O}_G to it. We have two types of queries. The first query, called a *degree query*, specifies a vertex v , and the oracle \mathcal{O}_G returns the degree of v . The second query, called a *neighbor query*, specifies a vertex v and an index i , and the oracle \mathcal{O}_G returns the i -th neighbor of v . A graph G is called ε -far from a property P if we must add or remove at least εm edges for it to satisfy the property P , where m is the number of edges. In

* JST, ERATO, Kawarabayashi Large Graph Project.

** Supported by JSPS Grant-in-Aid for Research Activity Start-up (24800082), MEXT Grant-in-Aid for Scientific Research on Innovative Areas (24106003), and JST, ERATO, Kawarabayashi Large Graph Project.

contrast to other models such as the adjacency matrix and the bounded-degree models, only a few properties are known to be efficiently testable in the adjacency list model. For examples, testing triangle-freeness, k -colorability for a constant k , and bipartiteness requires $\Omega(\sqrt{n})$ queries [2,3,9], where n is the number of vertices.

A graph G is called *isomorphic* to another graph H if there is a bijection $\pi : V(G) \rightarrow V(H)$ such that $(u, v) \in E(G)$ if and only if $(\pi(u), \pi(v)) \in E(H)$. In this paper, we consider the problem of testing if the input graph G is isomorphic to a fixed graph H , or if it is *H-isomorphic*. We assume that the (unknown) input graph G has the same number of vertices as H . The problem of deciding if a graph is isomorphic to H is fundamental and theoretically important. For example, the problem is one of the rare problems that is neither known to be in **P** nor **NP-Complete**. This motivates us to consider *H-isomorphism* in the property testing literature. A *graph property* refers to a property that is closed under taking isomorphism. Then, *H-isomorphism* can be identified as the simplest graph property such that every graph property can be expressed as a union of *H-isomorphisms*. Owing to these observations, Newman and Sohler [10] showed that every graph property is testable in the bounded-degree model if the input graph is a (bounded-degree) planar graph. This connection also holds for the adjacency list model, which motivates us to consider *H-isomorphism* in the adjacency list model.

If we assume that the input graph is an arbitrary graph possibly containing $\Omega(n^2)$ edges, testing *H-isomorphism* in the adjacency list model requires $\Omega(\sqrt{n})$ queries [4]. To investigate efficient testers for *H-isomorphism*, we restrict the input graph: We assume that the input graph and H are forests with the same number of vertices n . Note that we have no assumption on the degree as opposed to the bounded-degree model. To avoid uninteresting technicalities, we modify the definition of ε -farness as follows: Instead of using the number of edges in G to measure the distance, we say that a forest G is ε -far from isomorphic to a forest H if we must add or remove εn edges to transform G to H .³

With these definitions, we refer to the problem of testing the property of being isomorphic to a fixed forest as *testing forest-isomorphism*. The main result of this paper is as follows.

Theorem 1.1. *In the adjacency list model, we can test forest-isomorphism with $\text{polylog}(n)$ queries.*

Indeed, in our proof, we show that we can test forest-isomorphism even if both graphs are given as oracle accesses.

Further, we show a lower bound for testing forest-isomorphism.

Theorem 1.2. *In the adjacency list model, testing forest-isomorphism requires $\Omega(\sqrt{\log n})$ queries.*

³ Indeed, we often assume that the input graph contains $\Omega(n)$ edges in the adjacency list model. Thus, our definition of ε -farness for forests and the definition of ε -farness in the adjacency list model with the assumption are identical up to a constant multiplicative factor.

As a corollary of Theorem 1.1, we show the following general result.

Theorem 1.3. *In the adjacency list model, given an oracle access to a forest, we can test any graph property with $\text{polylog}(n)$ queries.*

Techniques We state a proof sketch of our main theorem, Theorem 1.1. Given a tree G , by removing εn edges from G , we can obtain a graph G' with the following property for some $s = s(\varepsilon)$. Each connected component of G' is either (i) a tree of maximum degree at most s , or (ii) a tree consisting of a (unique) root vertex of degree more than s and subtrees of size at most s .

The first step in our algorithm is providing an oracle access $\mathcal{O}_{G'}$ to G' using the oracle access \mathcal{O}_G to G . We call $\mathcal{O}_{G'}$ the *partitioning oracle*. In particular, if we specify a vertex v and an index i , the oracle $\mathcal{O}_{G'}$ returns whether the i -th edge incident to v in G is still alive in G' . By carefully designing the construction of G' , we can answer the query with $O(s^2)$ queries to \mathcal{O}_G .

Suppose that we have an oracle access $\mathcal{O}_{G'}$ to G' . Since we can deal with trees of type (i) using existing algorithms in the bounded-degree model, let us elaborate on trees of type (ii). For a tree T of type (ii), we can associate a tuple $(d, c_1, \dots, c_{t(s)})$ with it, where $t(s)$ is the number of possible trees of maximum degree at most s and size at most s . Note that $t(s)$ depends only on ε . Here, d is the degree of the root vertex of T , and c_i is the number of subtrees of the i -th type in T . Though we cannot exactly compute the tuple, given the root vertex of T , we can approximate it well using $\mathcal{O}_{G'}$. Since G' consists of trees of type (ii), we can associate a multiset of tuples with G' . We call it the *sketch* of G' . Though we cannot exactly compute the sketch, we can approximate it to some extent. The query complexity becomes $\text{polylog}(n)$ since we want to approximate d to within the multiplicative factor of $1 + \varepsilon$ and d can be up to n .

If G and H are isomorphic, then sketches associated with G' and H' must be the same. Our claim is that, if G' and H' are ε -far from being isomorphic, then their sketches are also far. Further, we will show that the distance between two sketches can be computed via maximum matching in the bipartite graph such that each vertex in the left part corresponds to a tree in G' and each vertex in the right part corresponds to a tree in H' . Since we can approximate sketches well and then approximate the size of the maximum matching from them, we obtain a tester for forest-isomorphism.

Related works There are two major models on the representation of graphs. In the *dense graph model*, a graph $G = (V, E)$ is given as an oracle $\mathcal{O}_G : V \times V \rightarrow \{0, 1\}$. Given two vertices $u, v \in V$, the oracle returns whether u and v are connected in G . A graph is called ε -far from a property P if we must add or remove at least εn^2 edges for it to satisfy P .

In the dense graph model, many properties such as triangle-freeness and k -colorability are known to be testable in constant time [6]. Indeed, Alon et al. [1] obtained the characterization of constant-time testable properties using Szemerédi's regularity lemma. As for graph isomorphism, Fischer and Matsliah [4] showed that testing H -isomorphism can be carried out with $\tilde{\Theta}(\sqrt{n})$ queries. If

both G and H are given as oracle accesses, then we need $\Omega(n)$ queries, and we can test with $\tilde{O}(n^{5/4})$ queries. We can trivially test forest-isomorphism: If a graph is isomorphic to a forest H , then it has at most n edges. If a graph is ε -far from being isomorphic to H , then it has at least $\varepsilon n^2 - n$ edges (otherwise, we can remove all edges and then add new edges to make H). Thus, we can distinguish the two cases only by estimating the number of edges up to, say $\frac{\varepsilon n^2}{2}$.

In the *bounded-degree model* with a degree bound d , a graph $G = (V, E)$ is given as an oracle $\mathcal{O}_G : V \times [d] \rightarrow V \cup \{\perp\}$, where $[d] = \{1, \dots, d\}$ and \perp is a special symbol. Given a vertex $v \in V$ and an index $i \in [d]$, the oracle returns the i -th neighbor of v . If there is no such neighbor, then the oracle returns \perp .

Many properties are known to be testable in constant time [7] and several general conditions of constant-time testability are shown [10,12]. Hassidim et al. [8] introduced the concept of the partitioning oracle to test minor closed properties. Our partitioning oracle is similar to theirs, but their oracle provides an oracle access to the graph that is determined by its internal random coin whereas ours provides an oracle access to a graph that is deterministically determined. As for graph isomorphism, it is known that H -isomorphism is testable in constant time when H is hyperfinite [10]. Here, a graph is *hyperfinite* if by removing εn edges, we can decompose the graph into connected components of size at most $f(\varepsilon)$ for some function f .

Organization In Section 2, we give notations and definitions used throughout the paper. In Section 3, we introduce the partitioning oracle. Using the partitioning oracle, it suffices to consider the case where each tree in the input graph is either a bounded-degree tree or a tree consisting of a high-degree root and subtrees of small sizes. In Section 4, we consider the case in which every tree in the input graph is the latter type and the degrees of roots are within a small interval. We deal with the general case and prove Theorem 1.1 in Section 5. Due to limitations of space, some proofs in Section 3, 4, 5 are presented in Appendix A, B, C. We prove Theorem 1.3 in Appendix D. We show the lower bound in Appendix E.

2 Preliminaries

For an integer n , we denote by $[n]$ the set $\{1, 2, \dots, n\}$ and denote by $\mathbb{N}_{<n}$ (resp. $\mathbb{N}_{\leq n}$) the set $\{0, 1, \dots, n-1\}$ (resp. $\{0, 1, \dots, n\}$).

Let $G = (V, E)$ be a graph. For a vertex v , $\deg_G(v)$ denotes the *degree* of v . We omit the subscript if it is clear from the context. For a set of vertices $S \subseteq V$, $G[S]$ denotes the subgraph *induced* by S . For graphs G and H with the same number of vertices, the *distance* $d(G, H)$ between G and H is defined as the minimum number of edges that need to be added or removed to make G isomorphic to H . Formally,

$$d(G, H) = \min_{\pi} (\#\{(u, v) \in E(G) \mid (\pi(u), \pi(v)) \notin E(H)\} \\ + \#\{(u, v) \notin E(G) \mid (\pi(u), \pi(v)) \in E(H)\}),$$

where π is over bijections from $V(G)$ to $V(H)$. We extend the definition of $d(G, H)$ for the case in which G and H have different number of vertices by adding a sufficient number of isolated vertices. For a graph G and an integer k , let $G + kv$ be the graph consisting of G and k isolated vertices. If $|V(G)| > |V(H)|$, we define $d(G, H) = d(G, H + (|V(G)| - |V(H)|)v)$. Similarly, if $|V(G)| < |V(H)|$, we define $d(G, H) = d(G + (|V(H)| - |V(G)|)v, H)$.

For an integer $s \geq 1$, we call a tree T an *s-rooted tree* if T contains a (unique) vertex v with $\deg(v) \geq s + 1$ such that each subtree of v contains at most s vertices. The vertex v is called the *root vertex* of T and is denoted by $\text{root}(T)$. We call a tree T an *s-bounded-degree tree* if every vertex in T has a degree of at most s . We call a tree T an *s-tree* if it is an *s-rooted tree* or an *s-bounded-degree tree*. To designate a union of trees, we use the term “forest.” For example, an *s-rooted forest* means a disjoint union of *s-rooted trees*.

3 Partitioning Oracle

In this section, we show that, for any $\varepsilon > 0$, there exists $s = s(\varepsilon)$ such that we can partition any forest into an *s-forest* by removing at most εn edges. Then, we show that we can provide an oracle access to the *s-forest*, which we call the *partitioning oracle*. We refer to a vertex with degree more than s in the original graph G as a *high-degree vertex*.

Lemma 3.1 (Partitioning oracle). *Suppose that we have an oracle access \mathcal{O}_G to a forest G in the adjacency list model. Then for every $\varepsilon > 0$, we can provide an oracle access \mathcal{O}'_G to a graph G' with the following properties:*

1. G' is an *s-forest* for some $s = s_{3.1}(\varepsilon)$. G' depends only on G and ε .
2. G' is obtained from G by removing at most εn edges.
3. Let V_h be high-degree vertices in G . Then, each tree in G' contains at most one vertex from V_h .

The oracle \mathcal{O}'_G supports alive-edge queries: Given a vertex v and an integer i , the oracle returns whether the i -th edge incident to v in G still exists in G' . For each alive-edge query, the oracle issues $O(1/\varepsilon^2)$ queries to \mathcal{O}_G . The output of \mathcal{O}'_G is deterministically calculated. Moreover, if G and H are isomorphic and $\Psi : V(G) \rightarrow V(H)$ is an isomorphism, $\mathcal{O}'_G(e) = \mathcal{O}'_H(\Psi(e))$ holds for every edge $e \in E(G)$.

Proof. We set $s = \frac{11}{\varepsilon}$. If the degree of a vertex is at most s , we call it *low-degree*. Let V_h and V_l be the sets of high-degree and low-degree vertices in G , respectively. We call a connected component in $G[V_l]$ *large* if it has more than s vertices and *small* otherwise. From the definition, there are at most $2n/s$ high-degree vertices in G and at most n/s large components in $G[V_l]$.

We first give a polynomial-time algorithm that outputs an *s-forest* from the input forest G . First, we remove edges (u, v) with $u, v \in V_h$ from G . Owing to this, the resulting graph can be seen as a bipartite graph, where the left part

is V_h and the right part consists of components in $G[V_l]$. Now for each small component C in $G[V_l]$, if it is adjacent to two or more vertices in V_h , we remove all the edges connecting C and V_h . Further, we remove all the edges between large components in $G[V_l]$ and V_h . We define G' as the resulting graph. As every subtree of each high-degree vertex is small, G' is an s -forest. Since each connected component of G' contains at most one high-degree vertex, the third property holds. Further, since any large small-degree connected component is not connected to a high-degree vertex, the first property holds. The total number of removed edges is at most $|V_h| + 2|V_h| + (n/s + 2|V_h|) = \varepsilon n$. Thus, the second property also holds.

We next show how to provide an oracle access to G' . We can support alive-edge queries as follows: Let $e = (v, w)$ be the queried edge. For v and w , we check if they are in V_h , in a large component of $G[V_l]$, or in a small component of $G[V_l]$. If they are in a small component of $G[V_l]$, we check whether the component is incident to two or more vertices in V_h . We can check these properties by performing a BFS in $G[V_l]$: If the BFS stops before visiting more than s vertices, it means that the vertex belongs to a small component. Otherwise, the vertex belongs to a large component. From this information, we can answer the alive-edge query. The total number of queries to \mathcal{O}_G is $O(s^2)$. From the argument above, answers to alive-edge queries are determined deterministically. \square

Since our construction of G' is deterministic and we remove at most εn edges, the following corollary holds.

Corollary 3.2. *Let G and H be two forests of n vertices, and G' and H' be the graphs obtained from G and H by the partitioning oracle with a parameter $\frac{\varepsilon}{4}$, respectively. If $d(G, H) = 0$, then $d(G', H') = 0$ holds. If $d(G, H) \geq \varepsilon n$, then $d(G', H') \geq \varepsilon n/2$ holds.* \square

Thus, we can preprocess the graph using the partitioning oracle, and it is sufficient to show that we can test isomorphism between two s -forests. We consider s -bounded-degree forests and s -rooted forests separately. Therefore, we construct a tester for the isomorphism of each corresponding tree in G' and H' . To test isomorphism between s -bounded-degree forests, we use a technique from [10]. We will develop a technique to test isomorphism between s -rooted forests in G' and H' under some conditions in the next section.

One technical issue of the partitioning oracle is that we cannot obtain the exact degree $\deg_{G'}(v)$ of a vertex v in G' since $\deg_G(v)$ can be up to n . Instead of computing the exact degree, we approximate the degree by randomly sampling incident edges as follows: Choose $i \in [\deg_G(v)]$ uniformly at random and apply the alive-edge query to the i -th incident edge. For a parameter $q \geq 1$, repeat this q times. Then, count the number of existing edges. Let c be this count. We use the value $\frac{c \deg_G(v)}{q}$ as an approximation to $\deg_{G'}(v)$ and denote it by $\widetilde{\deg}_{G',q}(v)$. The standard argument using Chernoff's bound gives the following lemma.

Lemma 3.3. *Let G' be the graph obtained from a graph G by the partitioning oracle. For any $\delta, \tau \in (0, 1)$ and a vertex v , there exists a polynomial $q = q_{3.3}(\delta, \tau)$ such that $\Pr[|\widetilde{\deg}_{G',q}(v) - \deg_{G'}(v)| \leq \delta \deg_G(v)] \geq 1 - \tau$.*

There is another issue of the partitioning oracle. If most parts of edges incident to a high-degree vertex v (i.e., a vertex with degree more than s) are removed by the partitioning oracle, the approximation $\deg_{G',q}(v)$ may have a considerably large relative error. However, we can ensure that the number of such high-degree vertices v is sufficiently small. To make the argument more formal, for an integer $R > s$, we call a vertex v *R-bad* if $R \cdot \max(\deg_{G'}(v), 1) \leq \deg_G(v)$. Otherwise, we call v *R-good*. Note that an *R-bad* vertex must satisfy $\deg_G(v) \geq R > s$. Thus, an *R-bad* vertex must be a high-degree vertex in G . Further, we call an s -rooted tree *R-bad* (resp. *R-good*) if the root vertex is *R-bad* (resp. *R-good*). Then, the number of vertices in *R-bad* s -rooted trees is bounded as follows.

Lemma 3.4. *Let G' be the s -forest obtained from a graph G by the partitioning oracle. For any $R > s$, the number of vertices in *R-bad* s -rooted trees of G' is at most $\frac{4sn}{R}$.*

Proof. Let B be the set of *R-bad* vertices and B' be the set of vertices in *R-bad* s -rooted trees. Since there are at most $2n/R$ vertices with $\deg_G(v) \geq R$, $|B| \leq 2n/R$ holds. From the third property of Lemma 3.1, each s -rooted tree in G' contains at most one high-degree vertex in G . Hence,

$$|B'| \leq \sum_{v \in B} (s \cdot \deg_{G'}(v) + 1) \leq \sum_{v \in B} \left(\frac{s \deg_G(v)}{R} + 1 \right) \leq \frac{4sn}{R}.$$

□

By Lemma 3.4, random vertex sampling does not pick up any *R-bad* vertex with high probability if R is chosen sufficiently large. In Section 4, assuming that every s -rooted tree is *R-good* in the input graph, we will construct a tester for forest-isomorphism. In Section 5, combining Lemma 3.4 and the tester given in Section 4, we will construct a tester for any s -forest.

For later use, we define auxiliary procedures on s -rooted trees. First, the following lemma is useful.

Lemma 3.5. *Given a vertex $v \in V(G')$ in an s -rooted tree T , there is an algorithm that finds a root vertex $\text{root}(T)$ with query complexity $O(\text{poly}(s))$.*

Proof. Perform a BFS in G' starting from the vertex v until we find a high-degree vertex. The third property of Lemma 3.1 guarantees that we can find the high-degree vertex and it is $\text{root}(T)$. □

Let $\mathcal{T}(s) = \{T^{(1)}, T^{(2)}, \dots, T^{(t(s))}\}$ be the family of all rooted trees with at most s vertices, where $t(s) = |\mathcal{T}(s)|$. For an s -rooted tree T , let $\text{Freq}(T)$ be the $t(s)$ -dimensional vector whose i -th coordinate is the number of subtrees of $\text{root}(T)$ isomorphic to $T^{(i)}$. As the root vertex uniquely exists in an s -rooted tree T , there is a unique $t(s)$ -dimensional vector corresponding to T .

Since the degree of a root vertex can be up to n , we cannot exactly compute $\text{Freq}(T)$. Instead, we approximate $\text{Freq}(T)$ by randomly sampling subtrees in T .

Given the root vertex v of an s -rooted tree T , we can define a procedure that approximates $\text{Freq}(T)$. We denote the procedure by $\widetilde{\text{Freq}}_q(v)$. The procedure $\widetilde{\text{Freq}}$ randomly samples an edge incident to v in G (rather than G') and invokes the alive-edge query. If the edge is alive, the procedure performs a BFS from the edge to obtain the whole subtree rooted at the edge. The procedure repeats this q times, where q is the parameter of the procedure. We give the procedure $\widetilde{\text{Freq}}$ in Algorithm 1. Again, Chernoff's bound guarantees the following.

Algorithm 1 Given the root vertex v of an s -rooted tree T and an integer q , the procedure $\widetilde{\text{Freq}}_q(v)$ returns an approximation to $\text{Freq}(T)$ by randomly sampling subtrees in T . The integer q represents the number of samples.

```

1: procedure  $\widetilde{\text{Freq}}_q(v)$ 
2:   Let  $\tilde{\mathbf{F}}$  be the all-zero  $t(s)$ -dimensional vector.
3:   for  $j = 1, \dots, q$  do
4:     Choose an integer  $k$  from  $[\deg_G(v)]$  uniformly at random.
5:     Ask whether the  $k$ -th edge  $(v, u)$  incident to  $v$  is alive.
6:     if the edge is alive then
7:       Perform a BFS from  $u$  to obtain the whole subtree rooted at  $u$ .
8:       Suppose that the subtree is isomorphic to  $T^{(i)}$ . Then, set  $\tilde{\mathbf{F}}[i] = \tilde{\mathbf{F}}[i] + 1$ .
9:   return  $(\deg_G(v)/q) \cdot \tilde{\mathbf{F}}$ 

```

Lemma 3.6. *For $s \geq 1$ and $\delta, \tau \in (0, 1)$, there exists a polynomial $q = q_{3.6}(s, \delta, \tau)$ such that for any s -rooted tree T , $|\text{Freq}(T)[i] - \widetilde{\text{Freq}}_q(\text{root}(T))[i]| \leq \delta \deg_G(v)$ for all $i \in [t(s)]$ with probability at least $1 - \tau$.*

Proof. Let $q_{3.6}(s, \delta, \tau) = O(\frac{\log(t(s)/\tau)}{\delta^2})$. By Chernoff's bound, it holds that $\Pr[|\text{Freq}(T)[i] - \widetilde{\text{Freq}}_{q_{3.6}}(v)[i]| > \delta \deg_G(v)] < \tau/t(s)$ for each i . By applying the union bound over all $i \in [t(s)]$, we obtain the lemma. \square

It is also useful to approximate the number of vertices in an s -rooted tree. For an s -rooted tree T , we can define a procedure $\widetilde{\text{Size}}$ that approximates $|V(T)|$ by randomly sampling the subtrees of T and computing the number of vertices in the subtrees. We give the procedure $\widetilde{\text{Size}}$ in Algorithm 2. the procedure $\widetilde{\text{Size}}$ first computes the approximate degree of the root vertex v of T by $\widetilde{\text{deg}}$ with sufficiently large samples. If $\widetilde{\text{deg}} = 0$, the procedure just returns 1 since T looks an isolated vertex. Otherwise, we randomly sample subtrees in G' q times, where q is the parameter of the procedure. For each subtree, we compute the number of vertices in the subtree. To randomly sample the subtrees, we randomly choose an edge in G (rather than G') until we choose an alive edge. This may take large amount of time since it is possible that most parts of edges incident to v are not alive. However, if T is guaranteed to be R -good for some $R > s$, the following holds.

Algorithm 2 Given two integers q, R and the root vertex v of an R -good s -rooted tree T , returns an approximation to $|V(T)|$ by randomly sampling the subtrees in T and computing the size of the subtrees. The integer q represents the number of samples.

```

1: procedure  $\widetilde{\text{Size}}_{G',q,R}(v)$ 
2:   Set  $q' = q_{3.3}(O(\delta/R), O(\tau))$  and compute  $\tilde{d} = \widetilde{\text{deg}}_{G',q'}(v)$ .
3:   if  $\text{deg}_G(v) < R$  then round  $\tilde{d}$  to the nearest integer.
4:   if  $\tilde{d} = 0$  then return 1
5:    $\tilde{S} = 0$ 
6:   for  $j = 1, \dots, q$  do
7:     loop
8:       Choose an integer  $k \in [\text{deg}_G(v)]$  uniformly at random.
9:       Ask whether the  $k$ -th edge  $(v, u)$  incident to  $v$  is alive.
10:      if the edge is alive then break
11:      Perform a BFS from  $u$  to obtain the size  $t$  of the subtree rooted at  $u$ .
12:       $\tilde{S} = \tilde{S} + t$ 
13:   return  $\tilde{d} \frac{\tilde{S}}{q} + 1$ 

```

Lemma 3.7. *For any $s, R \geq 1$ and $\delta, \tau \in (0, 1)$, there exists a polynomial $q = q_{3.7}(s, \delta, \tau)$ such that, for any R -good s -rooted tree T , $|\widetilde{\text{Size}}_{G',q,R}(\text{root}(T)) - |V(T)|| \leq \delta |V(T)|$ holds with probability at least $1 - \tau$. The expected number of queries issued by the procedure $\widetilde{\text{Size}}$ is $O(\text{poly}(s, R, \delta, \tau))$.*

The proof of Lemma 3.7 is a little complicated. We give the proof in Appendix A.

4 When All Root Vertices Have Similar Degrees

In this section and the next section, we assume that we read the input graphs G and H through the partitioning oracle. Thus, we are allowed to use alive-edge queries and the procedures deg , Freq , and Size . Further, we assume that s is a constant that depends only on ε .

We consider the case in which the root of all components have similar degrees. Formally, we assume that each component in G and H is R -good s -rooted tree and that the degree of each s -rooted tree in G and H is greater than B and at most γB . Here, $B(> s)$ is an integer that can be up to $O(n)$ and $s, \gamma \geq 1$ is an arbitrary constant. We call such a forest an *R -good s -rooted forest with root degrees in $(B, \gamma B]$* . In this section, we will show that there is a forest-isomorphism tester for R -good s -rooted forest with root degrees in $(B, \gamma B]$ whose query complexity is a polynomial in γ and R .

With the tester given in this section, we can construct a tester for the general case as follows. After applying the partitioning oracle, the graph becomes a disjoint union of an s -bounded-degree forest and an s -rooted forest. We partition the s -rooted forest into several groups by the root degree. First, we ignore all

the R -bad s -rooted trees from the graph. Since the number of R -bad trees is sufficiently small for a large R from Lemma 3.4, this does not affect so much. Second, if $\deg(\text{root}(T))$ is greater than $O(\gamma^i)$ and at most $O(\gamma^{i+1})$, we consider that a tree T is in the i -th group. Note that there are $O(\log n)$ groups. Then we apply the isomorphism tester of this section to each group. If input graphs G and H are isomorphic, the tester must return YES (isomorphic) for all the groups. In contrast, if G and H are ε -far from isomorphic, there must exist a group such that the tester returns NO (not isomorphic) for the group. Here, there is one technical issue: The number of vertices in such a group might be different.

We resolve this issue. We assume that $n := |V(G)|$ and $n' := |V(H)|$ might be slightly different and the algorithm does not know the exact values of n and n' but know their approximations. Formally, we assume that our algorithm will be given a value $\tilde{n} \geq 1$, an approximation to n and n' , and $\eta \in (0, 1)$ with $\frac{\tilde{n}}{n}, \frac{\tilde{n}}{n'} \in [1 - \eta, 1]$.

We can prove the following lemma.

Lemma 4.1. *Suppose that we are given $\varepsilon' > 0$, $\tilde{n} \geq 1$, $\gamma \geq 1$, $R, B > s$, $\tau \in (0, 1)$ and we can access s -forests G and H through the partitioning oracle, where $n = |V(G)|$ and $n' = |V(H)|$ might be different. Then, there exists $\eta = \eta_{4.1}(s, \varepsilon', \gamma, \tau, R) > 0$ with the following property. If G and H are R -good s -rooted forests with root degrees in $(B, \gamma B]$ with $\frac{\tilde{n}}{n}, \frac{\tilde{n}}{n'} \in [1 - \eta, 1]$, then there exists an algorithm that tests if $d(G, H) = 0$ or $d(G, H) \geq \varepsilon' \tilde{n}$ with probability at least $1 - \tau$. Assuming that s is constant, the query complexity is a polynomial in $R, \gamma, \varepsilon', \tau$ and does not depend on B, \tilde{n} . Further, denote by $q_{\text{random}}^{4.1}(s, \gamma, \varepsilon', \tau)$ the number of random vertex queries the algorithm invokes. Then, $q_{\text{random}}^{4.1}$ is a polynomial in $\gamma, \varepsilon', \tau$.*

In this section, we only write an overview of the proof of Lemma 4.1 since the proof is complicated, We provide the proof in Appendix B.

Since $\text{Freq}(T)$ maps to a unique $t(s)$ -dimensional vector corresponding to an s -rooted tree T , there is a unique multiset of vectors corresponding to an s -rooted forest G . For a $t(s)$ -dimensional vector $\mathbf{w} \in \mathbb{N}_{<n}^{t(s)}$, let $\Psi_G[\mathbf{w}]$ be the number of s -rooted trees T in G such that $\text{Freq}(T) = \mathbf{w}$. Note that Ψ_G can be seen as the sketch of G . Clearly, G is isomorphic to H if and only if $\Psi_G[\mathbf{w}] = \Psi_H[\mathbf{w}]$ for all \mathbf{w} . We use this property to create a tester. Since it is impossible to compute Ψ_G exactly, we resort to approximate it. We choose an integer $k \geq 1$, and divide each axis of the $t(s)$ -dimensional space into k segments to make $k^{t(s)}$ cells. We then estimate the number of s -rooted trees in each cell. We call this estimation the *sketch* of G . We focus on computing the sketch.

For an integer $k \geq 1$, we define intervals $I_i = [\frac{\tilde{n}i}{(1-\eta)k}, \frac{\tilde{n}(i+1)}{(1-\eta)k})$ ($i \in \mathbb{N}_{<k}$). Note that, for every $0 \leq i \leq n - 1$, there exists a unique interval I_j with $i \in I_j$. For a vector $\mathbf{u} \in \mathbb{N}_{<k}^{t(s)}$, let $\text{Cell}(\mathbf{u})$ be the corresponding cell formed by intervals $I_{\mathbf{u}[1]}, \dots, I_{\mathbf{u}[t(s)]}$. Further, for a vector $\mathbf{w} \in [0, n]^{t(s)}$, we define $\mathbf{Round}(\mathbf{w}) = \mathbf{u}$, where $\mathbf{u} \in \mathbb{N}_{<k}^{t(s)}$ is such that $\text{Cell}(\mathbf{u}) \ni \mathbf{w}$.

For a vector $\mathbf{u} \in \mathbb{N}_{<k}^{t(s)}$, we approximate the number of s -rooted trees T in G with $\text{Freq}(T) \in \text{Cell}(\mathbf{u})$ by the following algorithm $\widetilde{\text{Sketch}}$.

Algorithm 3 returns a map $\Phi : \mathbb{N}_{<k}^{t(s)} \rightarrow [0, n]$, given integers $q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k$, a real \tilde{n} and an R -good s -rooted forest G with root degrees in $(B, \gamma B]$ through the partitioning oracle. Here, $\Phi(\mathbf{u})$ is an approximation to the number of s -rooted trees T with $\text{Freq}(T) \in \text{Cell}(\mathbf{u})$.

```

1: procedure  $\widetilde{\text{Sketch}}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}(G)$ 
2:   Set  $\Phi(\mathbf{u}) = 0$  for all  $\mathbf{u} \in \mathbb{N}_{<k}^{t(s)}$ 
3:   for  $j = 1, \dots, q_{\text{loop}}$  do
4:     Choose a vertex  $u \in V(G)$  uniformly at random
5:     Perform a BFS from  $u$  to find a root vertex  $v$ .
6:      $\mathbf{u} = \text{Round}(\text{Freq}_{q_{\text{freq}}}(v))$ 
7:      $\Phi(\mathbf{u}) = \Phi(\mathbf{u}) + 1 / \widetilde{\text{Size}}_{G, q_{\text{size}}, R}(v)$ 
8:   return  $\frac{\tilde{n}}{q_{\text{loop}}} \Phi$ 

```

To create a forest-isomorphism tester, we first compute the sketches of G and H by the algorithm $\widetilde{\text{Sketch}}$, and then, we compute the minimum matching between the sketches. Here, the minimum matching is defined as the min-cost flow of complete bipartite graphs where vertices correspond to the cells of the sketches and the weight of an edge is the L1 distance between two cells of the sketches in the $t(s)$ -dimensional space. Since the L1 distance in the $t(s)$ -dimensional space corresponds to the number of different subtrees in s -rooted trees, we can prove that the a minimum matching between the sketches is a good approximation to $d(G, H)$ with high probability. Thus, it suffices to compute the sketches of G and H and the minimum matching between them. Note that we do not have to make any query to G and H to compute the minimum matching.

5 General Case

In this section, we prove Theorem 1.1. Missing parts of this section are given in Appendix C. Missing proofs are given in Appendix C.1. Again G and H denote the graphs given through the partitioning oracle and s is constant. For an integer $L \geq 1$, we call $G_1, \dots, G_L \subseteq G$ a *partition of G* if each G_i is a union of connected components in G and G is a disjoint union of G_1, \dots, G_L . The following lemma allows us to consider each part in the partition separately.

Lemma 5.1. *Let $L \geq 1$ be an integer and G_1, \dots, G_L (resp. H_1, \dots, H_L) be any partition of G (resp. H). Then, for any $\beta_1, \dots, \beta_L \geq 0$ summing up to 1, the following holds: For any $\varepsilon > 0$, if $d(G, H) \geq \varepsilon n$, there exists $i \in [L]$ such that $d(G_i, H_i) \geq \beta_i \varepsilon n$ holds.*

Proof. We can obtain the lemma immediately from the following claim.

Claim. $d(G, H) \leq \sum_{i=1}^L d(G_i, H_i)$.

We prove the claim. Construct a sequence of modifications to transform G to H . For each subgraph G_i with $|V(G_i)| \geq |V(H_i)|$, we transform G_i into H_i and $|V(G_i)| - |V(H_i)|$ isolated vertices. After this modification, for each subgraph G_i with $|V(G_i)| < |V(H_i)|$, we use G_i and $|V(H_i)| - |V(G_i)|$ isolated vertices to construct H_i . The total number of modifications is $\sum_i d(G_i, H_i)$. \square

To construct a tester for the isomorphism of s -forests, we first give a partition of an s -forest and apply Lemma 5.1. Then we test the isomorphism of each corresponding partition of G and H . That is, we check $d(G_i, H_i) = 0$ or $d(G_i, H_i) \geq \beta_i \varepsilon n$ for each i . Here, if $d(G, H) = 0$, all parts of the partition in G and H are isomorphic, so all the tests must output YES (with high probability). If $d(G, H) \geq \varepsilon n$, there must be an index i where the test outputs NO. To provide oracle accesses to G_i and H_i , we estimate the size of $V(G_i)$ and $V(H_i)$ by random sampling. If they are sufficiently far, we immediately return NO. If they are sufficiently small, we simply ignore G_i and H_i . Otherwise, we can provide the oracle accesses to G_i and H_i that costs for each query at most $\text{poly}(L)$ queries to G and H . Using this access, we test whether $d(G_i, H_i) \geq \beta_i \varepsilon n$.

To provide a partition of an s -forest, we introduce a new notion. For $\alpha, \gamma \geq 1$, $\mu > 0$, and a tree T , we say that T is *on the (α, γ, μ) -boundary*, if there exists an integer $i \geq 1$ with $1 - \mu \leq \deg(\text{root}(T)) / (\alpha \gamma^i) \leq 1 + \mu$. We denote by $B_{\alpha, \gamma, \mu}(G)$ the number of vertices in the trees of G that are on the (α, γ, μ) -boundary. For $\lambda > 0$, we call α *(γ, μ, λ) -good with respect to G* if $B_{\alpha, \gamma, \mu}(G) < \lambda n$. We can show that, if we choose α from $[1, \gamma]$ at random, α is (γ, μ, λ) -good with high probability.

Lemma 5.2. *Suppose that α is chosen from $[1, \gamma]$ uniformly at random. Then, for $\gamma \geq 2$, $\mu \in (0, 1/3)$, and $\lambda \in (0, 1)$, α is (γ, μ, λ) -good with respect to G with probability at least $1 - \frac{4\gamma\mu}{\lambda}$.*

We consider a partition of an s -forest G . Let α, γ, μ , and R be values chosen later. Let $G_{s, \alpha, \gamma, \mu, R}^{[0]}$ be the maximal s -bounded-degree forest in G and $G_{s, \alpha, \gamma, \mu, R}^{[1]}$ be the union of R -good s -rooted trees with root degree in $(s, \alpha\gamma]$ that are not on the (α, γ, μ) -boundary in G . Similarly, for $2 \leq i \leq L$, where $L = \lceil \log n / \log \gamma \rceil$, let $G_{s, \alpha, \gamma, \mu, R}^{[i]}$ be the union of R -good s -rooted trees with root degree in $(\alpha\gamma^{i-1}, \alpha\gamma^i]$ that are not on the (α, γ, μ) -boundary in G . Finally, let $G_{s, \alpha, \gamma, \mu, R}^{[L+1]}$ be the remaining trees that are not assigned to any partition so far. That is, $G^{[L+1]}$ is the union of trees that are R -bad or on the (α, γ, μ) -boundary in G . We omit the subscript of $G_{s, \alpha, \gamma, \mu, R}^{[i]}$ if it is clear from the context. Note that we can write $G = G^{[0]} \cup G^{[1]} \cup \dots \cup G^{[L+1]}$. We use the same notion for the other graph H .

We define a procedure that, given a vertex $v \in V(G)$, returns i with $v \in G^{[i]}$ as follows. Our procedure first determines if v is in an s -bounded-degree tree by performing a BFS from v until we visit $O(s)$ vertices. If we cannot find

a high-degree vertex, $v \in G^{[0]}$. Otherwise, for a parameter $q \geq 1$, we invoke $\widetilde{\deg}_q(\text{root}(v))$ and return an appropriate output. We call this procedure $\text{Which}_q(v)$.

Here, the technical issue is that the procedure Which may output a wrong value. We show that Which outputs the correct value with high probability for any partition of G except for $G^{[L+1]}$ and that the size of $G^{[L+1]}$ is sufficiently small.

Lemma 5.3. *For any $\tau \in (0, 1)$ and $R \geq 1$, there exists a polynomial $q = q_{5.3}(\gamma, \mu, R, \tau)$ such that the procedure $\text{Which}_q(v)$ outputs a correct value with probability $1 - \tau$ for $v \in V(G_{s, \alpha, \gamma, \mu, R}^{[0]}) \cup \dots \cup V(G_{s, \alpha, \gamma, \mu, R}^{[L]})$.*

Lemma 5.4. *For any $\gamma \geq 2$ and $\lambda \in (0, 1)$, there exist $R = O(s/\lambda)$, $\mu = O(\lambda/\gamma)$ such that if α is chosen from $[1, \gamma]$ uniformly at random, $|V(G_{s, \alpha, \gamma, \mu, R}^{[L+1]})| \leq \lambda n$ holds with probability $1 - O(1)$.*

Using the procedure Which , we can approximate the number of vertices in $G^{[i]}$ by random sampling. For $i \in \mathbb{N}_{\leq L}$, we denote by $\text{Size}_{q_{\text{loop}}, q_{\text{which}}}(G, i)$ the algorithm that samples q_{loop} vertices uniformly at random, and applies $\text{Which}_{q_{\text{which}}}$ for each sampled vertex, and then approximates $|V(G^{[i]})|$. By Chernoff's bound, we obtain the following lemma.

Lemma 5.5. *For any $\delta, \tau \in (0, 1)$ and parameters α, γ, μ , and R , there exist polynomials $q_{\text{loop}} = q_{\text{loop}5.5}(\delta, \tau)$ and $q_{\text{which}} = q_{\text{which}5.5}(\delta, \tau)$ such that the following holds: For any $\lambda \in (0, 1)$ with $|V(G_{s, \alpha, \gamma, \mu, R}^{[L+1]})| \leq \lambda n$, $|\text{Size}_{q_{\text{loop}}, q_{\text{which}}}(G, i) - |V(G_{s, \alpha, \gamma, \mu, R}^{[i]})|| \leq (\lambda + \delta)n$ with probability $1 - \tau$. \square*

Further, using the procedure Which , we can provide oracle accesses to $G^{[i]}$ for $i \in \mathbb{N}_{i \leq L}$. Let $\text{Random}_q(G, i)$ denote the procedure that repeats itself to pick up a vertex v in G uniformly at random and invokes the procedure $\text{Which}_q(v)$ and returns v if the returned value of Which is i .

Lemma 5.6. *For every $\delta, \tau \in (0, 1)$ and parameters α, γ, μ , and R , there exist polynomials $q = q_{5.6}(\delta, \tau)$ and $\lambda = \lambda_{5.6}(\delta, \tau)$ such that the following holds for every $i \in \mathbb{N}_{\leq L}$: If $|V(G^{[i]})| \geq \delta n$ and $|V(G^{[L+1]})| \leq \lambda n$, the procedure $\text{Random}_q(G, i)$ outputs a vertex of $G^{[i]}$ uniformly at random by invoking the procedure Which_q at most $O(1/(\delta\tau))$ times with probability $1 - \tau$.*

The sketch of the proof of Theorem 1.1 is as follows. As we mentioned, it suffices to create an isomorphism tester between $G^{[i]}$ and $H^{[i]}$ for each $i \in \mathbb{N}_{\leq L}$. First, set $\gamma = 2s$ and choose $\alpha \in [1, \gamma]$ uniformly at random. From Lemma 5.4, $|V(G^{[L+1]})|$ and $|V(H^{[L+1]})|$ are small with high probability. Thus, we can apply the procedures Which , Size and Random to the input graphs. From Lemmas 5.3, 5.5, and 5.6, these procedures output the correct value with sufficiently high probability. Using the procedure Size , we can test if $|V(G^{[i]})|$ and $|V(H^{[i]})|$ are large and sufficiently close. Then, we can test forest-isomorphism between $G^{[i]}$

and $H^{[i]}$ (with high probability) by providing oracle accesses to $G^{[i]}$ and $H^{[i]}$ through the procedure **Random**. For $i = 0$, we use a method proposed by [10] with a little modification. See Appendix C.2 for details. For $1 \leq i \leq L$, we use the method in Section 4. Here, every parameter depends on $\text{polylog}(n)$ assuming that s is constant. Thus, the query complexity of our forest-isomorphism tester is $\text{polylog}(n)$ in total. See Algorithm 5 in Appendix C.3 for the detailed description of the tester for forest-isomorphism.

References

1. N. Alon, E. Fischer, I. Newman, and A. Shapira. A combinatorial characterization of the testable graph properties: It’s all about regularity. *SIAM Journal on Computing*, 39(1):143–167, 2009.
2. N. Alon, T. Kaufman, M. Krivelevich, and D. Ron. Testing triangle-freeness in general graphs. *SIAM Journal on Discrete Mathematics*, 22(2):786–819, 2008.
3. I. Ben-Eliezer, T. Kaufman, M. Krivelevich, and D. Ron. Comparing the strength of query types in property testing: the case of testing k -colorability. In *SODA’08: Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1213–1222, 2008.
4. E. Fischer and A. Matsliah. Testing graph isomorphism. *SIAM Journal on Computing*, 38(1):207–225, 2008.
5. O. Goldreich. Introduction to testing graph properties. pages 105–141. Property Testing, 2010.
6. O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.
7. O. Goldreich and D. Ron. Property testing in bounded degree graphs. *Algorithmica*, 32(2):302–343, 2002.
8. A. Hassidim, J. A. Kelner, H. N. Nguyen, and K. Onak. Local graph partitions for approximation and testing. *FOCS’09: Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 22–31, 2009.
9. T. Kaufman, M. Krivelevich, and D. Ron. Tight bounds for testing bipartiteness in general graphs. *SIAM Journal on Computing*, 33(6):1441–1483, 2004.
10. I. Newman and C. Sohler. Every property of hyperfinite graphs is testable. *SIAM Journal on Computing*, 42(3):1095–1112, 2013.
11. R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.
12. S. Tanigawa and Y. Yoshida. Testing the supermodular-cut condition. *Algorithmica*, pages 1–11, 2013.
13. Y. Wu, Y. Yoshida, Y. Zhou, and A. Vijayraghavan. Graph isomorphism: Approximate and robust, 2013. manuscript.

A Proof of Lemma 3.7

Proof. First, we evaluate the probability that our procedure returns a good approximation. Let $S = |V(T)| - 1$ and set $q = q_{3.7} = O(\frac{s^2 \log(1/\tau)}{\delta^2})$. In Line 7–12, we choose an edge incident to v in G' uniformly at random. Therefore, by Chernoff’s bound, $|\frac{\tilde{S}}{q} - \frac{S}{\deg_{G'}(v)}| \leq O(\delta)$ holds with probability $1 - O(\tau)$. Since

we assume that T is R -good, at least one of (i) $R \deg_{G'}(v) > \deg_G(v)$ and (ii) $\deg_G(v) < R$ holds. To bound $|\widetilde{\deg_{G',q'}(v)} - S|$, let us consider these two cases.

When (i) holds but (ii) does not hold, v is not an isolated vertex in G' and \tilde{d} in the procedure is equal to the output of $\widetilde{\deg_{G',q'}(v)}$. From Lemma 3.3, $|\widetilde{\deg_{G',q'}(v)} - \deg_{G'}(v)| \leq O(\deg_G(v)/R)$ holds with probability $1 - O(\tau)$. The following claim is useful.

Claim. For any positive reals A, B, C, D with $|A - B| \leq \alpha$ and $|C - D| \leq \beta$, $|AC - BD| \leq \alpha D + \beta B + \alpha\beta$ holds.

Proof. By the triangle inequality,

$$\begin{aligned} |AC - BD| &= \frac{1}{2} |(A - B)(C + D) + (A + B)(C - D)| \\ &\leq \frac{1}{2} (|A - B||C + D| + |A + B||C - D|) \\ &\leq \frac{1}{2} (\alpha|C + D| + \beta|A + B|) \leq \alpha D + \beta B + \alpha\beta. \end{aligned}$$

□

From the claim and the union bound, we have $|\widetilde{\deg_{G',q'}(v)} - S| \leq O(\delta \deg_{G'}(v)) + O(\frac{\delta \deg_G(v)}{R \deg_{G'}(v)} S) + O(\delta \deg_G(v)/R) \leq O(\delta S) + O(\delta S) + O(\delta S) = \delta S$ with probability $1 - \tau$.

When (ii) holds, $\tilde{d} = \deg_{G'}(v)$ holds (with probability $1 - O(\tau)$) by rounding in Line 3. Thus, if v is an isolated vertex in G' , our procedure will return 1 in Line 4. Otherwise, $|\tilde{d} \frac{\tilde{S}}{q} - S| = |\frac{\tilde{S}}{q} - \frac{S}{\deg_{G'}(v)}| \cdot \tilde{d} \leq \delta \tilde{d} \leq \delta |V(T)|$ holds with probability $1 - \tau$.

Next, we turn to analyze the expected number of queries issued by the procedure. Since q' is $\text{poly}(R, \delta, \tau)$, we make at most $\text{poly}(R, \delta, \tau)$ queries to compute $\widetilde{\deg_{G',q'}(v)}$ in Line 2. Further, since we assume that T is R -good, Line 7–12 takes $O(R + \text{poly}(s))$ time on average. Thus, the expected query complexity is $O(\text{poly}(s, R, \delta, \tau))$ in total. □

B Proof of Lemma 4.1

In this section, we prove Lemma 4.1. We use the notions defined in Section 4. Throughout this section, $c(G)$ denotes the number of connected components in G .

This section is organized as follows. First, we analyze the behavior of the algorithm **Sketch** in Algorithm 3 in Section B.1. Next, we discuss the formal definition of the minimum matching between sketches in Section B.2. Finally, we show that the minimum matching is a good approximation to $d(G, H)$ and we give a tester for isomorphism in Section B.3.

B.1 Approximation algorithms for sketches

In this subsection, we analyze the behavior of the algorithm $\widetilde{\text{Sketch}}$. Upon computing the sketch, it is desired that the procedure $\widetilde{\text{Size}}_{G, q_{\text{size}}, R}(T)$ always outputs a good approximation to $|V(T)|$ for an s -rooted tree T . For $\delta' \in (0, 1)$ and an s -rooted tree T , we say that (the output of) $\widetilde{\text{Size}}_{G, q_{\text{size}}, R}(T)$ is δ' -safe if $|\widetilde{\text{Size}}_{G, q_{\text{size}}, R}(\text{root}(T)) - |V(T)|| \leq \delta' |V(T)|$ holds. Further, we say that (the execution of) $\widetilde{\text{Sketch}}$ is δ' -safe if all the outputs of $\widetilde{\text{Size}}$ in Line 7 are δ' -safe. Let $\text{Sketch}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}^{\delta'}(G)(\mathbf{u}) = \mathbf{E}[\widetilde{\text{Sketch}}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}(G)(\mathbf{u}) \mid \widetilde{\text{Sketch}} : \delta'\text{-safe}]$. Note that $\widetilde{\text{Sketch}}$ is δ' -safe with high probability if the parameters are chosen appropriately by Lemma 3.7. Let $G^{(i)}$ ($i \in [c(G)]$) be the i -th s -rooted tree in G and $v^{(i)} = \text{root}(G^{(i)})$. We use the following two lemmas in the next subsection.

Lemma B.1. *For any $k, s, R, B \geq 1$ and $\delta, \delta', \tau \in (0, 1)$, there exist $q_{\text{loop}} = q_{\text{loop}}^{B.1}(k, s, \delta, \delta', \tau)$ and $q_{\text{size}} = q_{\text{size}}^{B.1}(k, s, \delta, \delta', \tau)$ such that for any q_{freq} and an R -good s -rooted forest G with root degrees in $(B, \gamma B]$, $|\widetilde{\text{Sketch}}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}(G)(\mathbf{u}) - \text{Sketch}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}^{\delta'}(G)(\mathbf{u})| \leq \frac{\delta n}{B}$ holds for all $\mathbf{u} \in \mathbb{N}_{< k}^{t(s)}$ with probability at least $1 - \tau$. Here $q_{\text{loop}}^{B.1}$ and $q_{\text{size}}^{B.1}$ are polynomials in $k^{t(s)}, \delta, \delta', \tau$.*

Proof. For simplicity, we omit the subscript of procedures. Let $p_{i, \mathbf{u}}$ be the probability that a vertex of $G^{(i)}$ is chosen in Line 4 and \mathbf{u} is obtained in Line 6 of $\widetilde{\text{Sketch}}$. Denote by Φ the mapping in the algorithm $\widetilde{\text{Sketch}}$.

For $\delta' \in (0, 1)$, it holds that

$$\begin{aligned} \text{Var}[\Phi(\mathbf{u}) \mid \widetilde{\text{Sketch}} : \delta'\text{-safe}] &\leq \sum_{i \in [c(G)]} p_{i, \mathbf{u}} \cdot \mathbf{E} \left[\frac{1}{\widetilde{\text{Size}}_{G, q_{\text{size}}, R}(v^{(i)})^2} \mid \widetilde{\text{Size}} : \delta'\text{-safe} \right] \\ &\leq \sum_{i \in [c(G)]} \frac{|V(G^{(i)})|}{n} \cdot \left(\frac{1}{(1 - \delta')|V(G^{(i)})|} \right)^2. \end{aligned}$$

Further, since we assume that the root degree of each s -tree in G is in $(B, \gamma B]$, $c(G) \leq n/B$ holds. By Chebyshev's inequality,

$$\begin{aligned}
& \Pr \left[\left| \widetilde{\text{Sketch}}(\mathbf{u}) - \text{Sketch}^{\delta'}(\mathbf{u}) \right| \geq \frac{\delta n}{B} \mid \widetilde{\text{Sketch}} : \delta'\text{-safe} \right] \\
& \leq \left(\frac{B}{\delta n} \right)^2 \cdot \left(\frac{\tilde{n}}{q_{\text{loop}}} \right)^2 \cdot \mathbf{Var}[\Phi(\mathbf{u}) \mid \widetilde{\text{Sketch}} : \delta'\text{-safe}] \\
& \leq \left(\frac{B}{\delta n} \right)^2 \cdot \left(\frac{\tilde{n}}{q_{\text{loop}}} \right)^2 \cdot q_{\text{loop}} \sum_{i \in [c(G)]} \frac{|V(G^{(i)})|}{n} \cdot \left(\frac{1}{(1-\delta')|V(G^{(i)})|} \right)^2 \\
& \leq \frac{1}{q_{\text{loop}} \delta^2 (1-\delta')^2} \sum_{i \in [c(G)]} \frac{B^2}{n |V(G^{(i)})|} \\
& \leq \frac{1}{q_{\text{loop}} \delta^2 (1-\delta')^2}.
\end{aligned}$$

Here, in the last inequality, note that

$$\sum_{i \in [c(G)]} \frac{B^2}{n |V(G^{(i)})|} \leq \sum_{i \in [c(G)]} \frac{B}{n} = \frac{c(G)B}{n} \leq 1.$$

Set $q_{\text{loop}}^{B.1}(s, k, \delta, \delta', \tau) = O(k^{t(s)} / (\delta^2 (1-\delta')^2 \tau))$ and $q_{\text{size}}^{B.1}(s, k, \delta, \delta', \tau) = q_{3.7}(s, \delta', \tau / (2q_{\text{loop}}))$.

Then, the execution of $\widetilde{\text{Sketch}}$ is δ' -safe with probability $1 - \tau / (2k^{t(s)})$. Therefore, concerning the conditional probability, $|\widetilde{\text{Sketch}}(\mathbf{u}) - \text{Sketch}^{\delta'}(\mathbf{u})| < \frac{\delta n}{B}$ holds with probability $1 - \tau / k^{t(s)}$. Applying the union bound for all $\mathbf{u} \in \mathbb{N}_{<k}^{t(s)}$, the lemma follows. \square

Lemma B.2. *For any $s \geq 1$ and $\delta \in (0, 1)$, there exist linear functions $\delta' = \delta'_{B.2}(\delta)$ and $\eta = \eta_{B.2}(\delta)$ such that if $\frac{\tilde{n}}{n} \in [1-\eta, 1]$, then for any $R, k, B, \gamma, q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}} \geq 1$ and R -good s -rooted forest G with root degrees in $(B, \gamma B]$, $\|\text{Sketch}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}^{\delta'}(G)\|_{1-c(G)} \leq \delta c(G)$ holds.*

Proof. Again, for simplicity, we omit the subscript of procedures. For $u \in V(G)$, let $v(u)$ be a root vertex of an s -rooted tree that u belongs to.

$$\begin{aligned}
\|\text{Sketch}^{\delta'}(G)\|_1 &= \frac{\tilde{n}}{q_{\text{loop}}} \cdot q_{\text{loop}} \cdot \mathbf{E}_{u \in V(G)} \left[\frac{1}{\widetilde{\text{Size}}_{G, q_{\text{size}}, R}(v(u))} \mid \widetilde{\text{Size}} : \delta'\text{-safe} \right] \\
&= \tilde{n} \sum_{i \in [c(G)]} \Pr[\text{vertex of } G^{(i)} \text{ is chosen}] \cdot \mathbf{E} \left[\frac{1}{\widetilde{\text{Size}}_{G, q_{\text{size}}, R}(v^{(i)})} \mid \widetilde{\text{Size}} : \delta'\text{-safe} \right] \\
&= \tilde{n} \sum_{i \in [c(G)]} \frac{|V(G^{(i)})|}{n} \cdot \mathbf{E} \left[\frac{1}{\widetilde{\text{Size}}_{G, q_{\text{size}}, R}(v^{(i)})} \mid \widetilde{\text{Size}} : \delta'\text{-safe} \right].
\end{aligned}$$

By the condition of the lemma, $\tilde{n}/n \in [1 - \eta, 1]$. Further, the expectation in the last equation is between $1/((1 + \delta')|V(G^{(i)})|)$ and $1/((1 - \delta')|V(G^{(i)})|)$. Therefore, $\|\text{Sketch}^{\delta'}(G)\|_1 \in [(1 - \eta)c(G)/(1 + \delta'), c(G)/(1 - \delta')]$. Setting $\eta_{B.2}(\delta) = \delta'_{B.2}(\delta) = \delta/2$, the lemma follows. \square

Hereafter, for $\delta, \tau \in (0, 1)$ and $q_{\text{freq}}, k, R \geq 1$, we denote by $\widetilde{\text{Sketch}}_{\delta, \tau, q_{\text{freq}}, R, k}$ the procedure $\widetilde{\text{Sketch}}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}$ for $\delta' = \delta'_{B.2}(\delta)$, $q_{\text{loop}} = q_{\text{loop}}^{B.1}(k, s, \delta, \delta', \tau)$, and $q_{\text{size}} = q_{\text{size}}^{B.1}(k, s, \delta, \delta', \tau)$ in order to simplify the notion. Further, we denote by $\text{Sketch}_{\delta, \tau, q_{\text{freq}}, R, k}$ the conditional expectation $\text{Sketch}_{q_{\text{loop}}, q_{\text{freq}}, q_{\text{size}}, R, k}^{\delta'}$.

Finally, we consider the (expected) query complexity of $\widetilde{\text{Sketch}}$. Since $q_{\text{loop}}^{B.1}, q_{\text{size}}^{B.1}$ are polynomials in k, δ , and τ , the following holds.

Lemma B.3. *For $\delta, \tau \in (0, 1)$ and $R, k \geq 1$, the expected query complexity of the algorithm $\widetilde{\text{Sketch}}_{\delta, \tau, q_{\text{freq}}, R, k}$ is a polynomial in $\delta, \tau, q_{\text{freq}}, R, k^{t(s)}$.* \square

B.2 Matching sketches

In this subsection, we define the distance between two sketches so that it is a good approximation to $d(G, H)$. Let denote by $\mathbb{R}_{\geq 0}$ the set of non-negative reals.

A *weighted point set* is a tuple $X = (w, S)$, where $w : S \rightarrow \mathbb{R}_{\geq 0}$ is a *weight function* and S is a set of vectors. To define the distance between sketches, we consider the following problem.

Definition B.4 (Minimum matching between weighted point sets). *Let $X_1 = (w_1, S_1)$ and $X_2 = (w_2, S_2)$ be two weighted points sets with $\|w_1\|_1 = \|w_2\|_1$. We call a function $f : S_1 \times S_2 \rightarrow \mathbb{R}_{\geq 0}$ a flow function from X_1 to X_2 if $\sum_{\mathbf{v} \in S_2} f(\mathbf{u}, \mathbf{v}) = w_1(\mathbf{u})$ for all $\mathbf{u} \in S_1$ and $\sum_{\mathbf{u} \in S_1} f(\mathbf{u}, \mathbf{v}) = w_2(\mathbf{v})$ for all $\mathbf{v} \in S_2$. The value of a flow function f is defined as*

$$\sum_{(\mathbf{u}, \mathbf{v}) \in S_1 \times S_2} f(\mathbf{u}, \mathbf{v}) \cdot \|\mathbf{u} - \mathbf{v}\|_1.$$

The minimum value of a flow function is denoted by $\mathcal{M}(X_1, X_2)$, and the flow function that achieves the minimum value is called the optimal flow function.

Note that the optimal flow function can be calculated by a min-cost flow algorithm on a bipartite graph. Therefore, the following lemma holds.

Lemma B.5. *Let $X_1 = (w_1, S_1)$ and $X_2 = (w_2, S_2)$ be weighted point sets. If w_1 and w_2 are integral, there exists an optimal flow function f^* that is integral. In particular, if all values of w_1 and w_2 are 1, the set of pairs $\{(\mathbf{u}, \mathbf{v}) \in S_1 \times S_2 \mid f^*(\mathbf{u}, \mathbf{v}) = 1\}$ forms a matching.* \square

For an s -rooted forest G , let $\text{Freq}(G)$ denote the multiset $\{\text{Freq}(G^{(1)}), \dots, \text{Freq}(G^{(c(G))})\}$. To define the distance between sketches, we first associate weighted point sets F_G , S_G , and \tilde{S}_G with $\text{Freq}(G)$, $\text{Sketch}(G)$, and $\widetilde{\text{Sketch}}(G)$, respectively. Then,

we show that $\mathcal{M}(F_G, F_H)$ can be well approximated by $\mathcal{M}(\tilde{S}_G, \tilde{S}_H)$. Next, we show that $d(G, H)$ can be approximated by $\mathcal{M}(F_G, F_H)$. Since we can efficiently compute \tilde{S}_G and \tilde{S}_H , it follows that we can well approximate $d(G, H)$. Hence, we can test isomorphism between G and H .

We first introduce auxiliary weighted point sets. Let $F'_G = (\mathbf{1}, \text{Freq}(G))$, where $\mathbf{1}$ is the constant-one function. For parameters $\delta, \tau, q_{\text{freq}}, R, k$, we define $S'_{G, \delta, \tau, q_{\text{freq}}, R, k}$ as follows. First for a cell C , we define $\mathbf{vtx}(C)$ as the unique point in C that is minimal with respect to every axis. Then, for each $\mathbf{u} \in \mathbb{N}_{<k}^{t(s)}$, we add a point $\mathbf{vtx}(\text{Cell}(\mathbf{u}))$ with weight $\text{Sketch}_{\delta, \tau, q_{\text{freq}}, R, k}(G)(\mathbf{u})$. Similarly, we define $\tilde{S}'_{G, \delta, \tau, q_{\text{freq}}, R, k}$. If the parameters are clear from the context, we occasionally drop the subscripts of \tilde{S}' and S' . A technical issue here is that the sums of weights of F'_G , S'_G , and \tilde{S}'_G might be different since Sketch is a random variable, and it means that we cannot define matchings among them. To avoid this issue, for a large integer value M , we define $\text{ext}((w', S'), M) = (w, S)$ as the extension of (w', S') so that $S = S' \cup \{\perp\}$ and $w(\perp) = M - \|w'\|_1$. We regard \perp as the all-zero vector when measuring distances to other vectors. For a sufficiently large M , we define $F_G = \text{ext}(F'_G, M)$, $S_G = \text{ext}(S'_G, M)$, and $\tilde{S}_G = \text{ext}(\tilde{S}'_G, M)$.

This section is devoted to prove the following lemma.

Lemma B.6. *For any $s, R, \gamma \geq 1$, and $\delta'', \tau' \in (0, 1)$, there exist parameters $\delta, \tau, q_{\text{freq}}$, and k such that $|\mathcal{M}(F_G, F_H) - \mathcal{M}(\tilde{S}_G, \tilde{S}_H)| \leq \delta''n$ holds with probability at least $1 - \tau'$. The parameters δ, k are polynomials in γ, δ'' , τ is $O(\tau')$, and q_{freq} is a polynomial in γ, δ'', R .*

To prove Lemma B.6, we prove several lemmas first.

Lemma B.7. *For any $s, \gamma, q_{\text{freq}} \geq 1$, and $\delta'', \tau \in (0, 1)$, there exists $\delta = \delta_{B.7}(s, \gamma, \delta'')$ such that*

$$\mathcal{M}(S_{G, \delta, \tau, q_{\text{freq}}, R, k}, \tilde{S}_{G, \delta, \tau, q_{\text{freq}}, R, k}) \leq \delta''n$$

with probability at least $1 - \tau$. Here $\delta_{B.7}$ is a polynomial in γ, δ'' .

Proof. We construct a flow function from S_G to \tilde{S}_G so that $\sum_{(\mathbf{u}, \mathbf{v})} f(\mathbf{u}, \mathbf{v}) \|\mathbf{u} - \mathbf{v}\|_1 \leq \delta''n$ holds with high probability. We assign $f(\mathbf{u}, \mathbf{u}) = \min(\text{Sketch}(G)(\mathbf{u}), \widetilde{\text{Sketch}}(G)(\mathbf{u}))$ for each $\mathbf{u} \in \mathbb{N}_{<k}^{t(s)}$ and assign an arbitrary value to other parts of f so that f satisfies the condition of a flow function. By Lemma B.1, $\sum_{\mathbf{u}, \mathbf{v} \in \mathbb{N}_{<k}^{t(s)}, \mathbf{u} \neq \mathbf{v}} f(\mathbf{u}, \mathbf{v}) \leq \frac{\delta n}{B}$ with probability at least $1 - \tau$. Set $\delta = \delta_{B.7}(s, \gamma, \delta'') = \delta''/(t(s)\gamma)$. Since $\|\mathbf{u} - \mathbf{v}\|_1 \leq t(s)\gamma B$, we have

$$\sum_{(\mathbf{u}, \mathbf{v})} f(\mathbf{u}, \mathbf{v}) \|\mathbf{u} - \mathbf{v}\|_1 \leq \frac{\delta n}{B} \cdot t(s)\gamma B \leq \delta''n.$$

□

Lemma B.8. *For any $s, R, \gamma \geq 1$ and $\delta'', \tau \in (0, 1)$, there exist $k = k_{B.8}(s, \gamma, \delta'')$, $q_{\text{freq}} = q_{\text{freq} B.8}(s, \gamma, \delta'', R)$, $\delta = \delta_{B.8}(s, \gamma, \delta'')$ such that $\mathcal{M}(F_G, S_{G, \delta, \tau, q_{\text{freq}}, R, k}) \leq \delta''n$ holds. The parameters $k_{B.8}, \delta_{B.8}$ are polynomials in γ, δ'' and $q_{\text{freq} B.8}$ is a polynomial in γ, δ'', R .*

Proof. Let $k, \delta, q_{\text{freq}}$ be parameters chosen later. Again, we construct a flow function f from F_G to S_G . We define a hypercube C_i in $\mathbb{R}^{t(s)}$ as $C_i = \{(x_1, x_2, \dots, x_{t(s)}) \mid |x_j - \text{Freq}(G^{(i)})[j]| < \gamma B/(2k), j \in [t(s)]\}$. Further, let $B_i = \{\mathbf{v} \mid \text{Cell}(\mathbf{v}) \cap C_i \neq \emptyset\}$. Note that $|B_i| \leq 2^{t(s)}$.

Let $\delta' = \delta'_{B.2}(\delta)$. Let $p^{(i)}$ be the probability that a vertex of $G^{(i)}$ is chosen in Line 4 of the algorithm $\widetilde{\text{Sketch}}$, $q_{i,\mathbf{v}}$ be the probability that \mathbf{v} holds, and $e^{(i)} = \mathbf{E} \left[1/\widetilde{\text{Size}}(v^{(i)}) \mid \widetilde{\text{Size}} : \delta'\text{-good} \right]$. For $i \in [c(G)]$ and $\mathbf{v} \in \mathbb{N}_{<k}^{t(s)}$, define a flow function as follows.

$$f(i, \mathbf{v}) = (1 - \delta') \tilde{n} p^{(i)} q_{i,\mathbf{v}} e^{(i)}$$

We show that $\sum_i f(i, \mathbf{v}) \leq \text{Sketch}(G)(\mathbf{v})$ for all \mathbf{v} and $\sum_{\mathbf{v}} f(i, \mathbf{v}) \leq 1$ for all i . The conditional expectation Sketch can be expressed as $\text{Sketch}(G)(\mathbf{v}) = \sum_i \tilde{n} p^{(i)} q_{i,\mathbf{v}} e^{(i)}$. Thus, $\sum_i f(i, \mathbf{v}) = (1 - \delta') \text{Sketch}(G)(\mathbf{v}) < \text{Sketch}(G)(\mathbf{v})$ holds. Further, since $p^{(i)} e^{(i)} \leq (|V(G^{(i)})|/n) \cdot 1/((1 - \delta')|V(G^{(i)})|) = 1/(n(1 - \delta'))$ and $\sum_{\mathbf{v}} q_{i,\mathbf{v}} = 1$, $\sum_{\mathbf{v}} f(i, \mathbf{v}) \leq 1$ holds. Similarly, we can show that $\sum_{\mathbf{v}} f(i, \mathbf{v}) \geq (1 - \delta')/(1 + \delta') \geq 1 - 2\delta'$ for all i .

We assign values to the remaining part of f so that the condition of the flow function is satisfied. Here it holds that $\sum_{\mathbf{v}} f(\perp, \mathbf{v}) = \sum_{\mathbf{v}} (\text{Sketch}(G)(\mathbf{v}) - \sum_i f(i, \mathbf{v})) = \delta' \|\text{Sketch}(G)\|_1$ and $\sum_i f(i, \perp) = \sum_i (1 - \sum_{\mathbf{v}} f(i, \mathbf{v})) \leq 2\delta' c(G)$. From Lemma B.2, we have $\sum_{\mathbf{v}} f(\perp, \mathbf{v}) + \sum_i f(i, \perp) \leq 4\delta' c(G)$.

Let $r_i = \sum_{\mathbf{v} \in B_i} q_{i,\mathbf{v}}$. Set $k = k_{B.8}(s, \gamma, \delta'') = O(2^{t(s)} t(s) \gamma / \delta'')$ and $q_{\text{freq}} = q_{\text{freq}B.8}(s, \gamma, \delta'', R) = q_{3.6}(s, 1/(2kR), \tau')$ for $\tau' = O(1/(k^{t(s)} \cdot t(s) \gamma))$. Then from Lemma 3.6, we have $r_i \geq 1 - \tau'$.

Now, we calculate the value of the flow function. For fixed $i \in [c(G)]$,

$$\begin{aligned} \sum_{\mathbf{v}} f(i, \mathbf{v}) \cdot \|\text{Freq}(G^{(i)}) - \mathbf{v}\|_1 &= \sum_{\mathbf{v} \in B_i} f(i, \mathbf{v}) \cdot \|\text{Freq}(G^{(i)}) - \mathbf{v}\|_1 + \sum_{\mathbf{v} \notin B_i} f(i, \mathbf{v}) \cdot \|\text{Freq}(G^{(i)}) - \mathbf{v}\|_1 \\ &\leq \sum_{\mathbf{v} \in B_i} 1 \cdot \frac{t(s) \gamma B}{k} + \sum_{\mathbf{v} \notin B_i} (1 - r_i) \cdot t(s) \gamma B \\ &\leq 2^{t(s)} \cdot 1 \cdot \frac{t(s) \gamma B}{k} + k^{t(s)} \tau' \cdot t(s) \gamma B \\ &\leq \frac{\delta'' B}{4} + \frac{\delta'' B}{4} = \frac{\delta'' B}{2}. \end{aligned}$$

Set $\delta = \delta_{B.8}(s, \gamma, \delta'') = O(\delta''/(k^{t(s)} \gamma))$ so that $\delta' = O(1/(k^{t(s)} \gamma))$. Then we have

$$\begin{aligned} \sum_{i, \mathbf{v}} f(i, \mathbf{v}) \cdot \|\text{Freq}(G^{(i)}) - \mathbf{v}\|_1 &\leq c(G) \cdot \frac{\delta'' B}{2} = \frac{\delta'' c(G) B}{2} \\ \sum_{\mathbf{v}} f(\perp, \mathbf{v}) \|\mathbf{v}\|_1 + \sum_i f(\text{Freq}(G^{(i)}), \perp) \|\text{Freq}(G^{(i)})\|_1 &\leq 4\delta' c(G) \cdot k^{t(s)} \gamma B = \frac{\delta'' c(G) B}{2}. \end{aligned}$$

Since $c(G)B \leq n$, the cost of the flow function is at most $\delta'' n$. \square

We can show that the triangle inequality holds for the minimum value of a flow function \mathcal{M} . Combining Lemmas B.7, B.8, and the triangle inequality, we can prove Lemma B.6.

Lemma B.9. *Let $X_i = (w_i, S_i)$ ($i = 1, 2, 3$) be weighted point sets with $\|w_1\|_1 = \|w_2\|_1 = \|w_3\|_1$. Then, the triangle inequality holds for $\mathcal{M}(\cdot, \cdot)$ among them, that is*

$$\mathcal{M}(X_1, X_3) \leq \mathcal{M}(X_1, X_2) + \mathcal{M}(X_2, X_3).$$

Proof. We construct a flow function f_{13} from X_1 to X_3 as follows. Let f_{12}^* be the optimal flow function from X_1 to X_2 , and let f_{23}^* be the one from X_2 to X_3 . Let $f_{13}(i_1, i_3) = \sum_{i_2} \frac{f_{12}^*(i_1, i_2) f_{23}^*(i_2, i_3)}{w_2(i_2)}$. The function f_{13} satisfies the conditions of a flow function:

$$\begin{aligned} \sum_{i_3} f_{13}(i_1, i_3) &= \sum_{i_3} \sum_{i_2} \frac{f_{12}^*(i_1, i_2) f_{23}^*(i_2, i_3)}{w_2(i_2)} = \sum_{i_2} f_{12}^*(i_1, i_2) = w_1(i_1), \\ \sum_{i_1} f_{13}(i_1, i_3) &= \sum_{i_1} \sum_{i_2} \frac{f_{12}^*(i_1, i_2) f_{23}^*(i_2, i_3)}{w_2(i_2)} = \sum_{i_2} f_{23}^*(i_2, i_3) = w_3(i_3). \end{aligned}$$

We observe that

$$\begin{aligned} &\sum_{i_1, i_3} f_{13}(i_1, i_3) \|S_1(i_1) - S_3(i_3)\|_1 \\ &= \sum_{i_1, i_3} \sum_{i_2} \frac{f_{12}^*(i_1, i_2) f_{23}^*(i_2, i_3) \|S_1(i_1) - S_3(i_3)\|_1}{w_2(i_2)} \\ &\leq \sum_{i_1, i_3} \sum_{i_2} \frac{f_{12}^*(i_1, i_2) f_{23}^*(i_2, i_3) (\|S_1(i_1) - S_2(i_2)\|_1 + \|S_2(i_2) - S_3(i_3)\|_1)}{w_2(i_2)} \\ &= \sum_{i_1, i_2} f_{12}^*(i_1, i_2) \|S_1(i_1) - S_2(i_2)\|_1 + \sum_{i_2, i_3} f_{23}^*(i_2, i_3) \|S_2(i_2) - S_3(i_3)\|_1. \end{aligned}$$

Thus, $\mathcal{M}(X_1, X_3) \leq \mathcal{M}(X_1, X_2) + \mathcal{M}(X_2, X_3)$. \square

Proof (Proof of Lemma B.6). Set $q_{\text{freq}} = q_{\text{freq}B.8}(s, \gamma, O(\delta''), R)$, $\delta = \min(\delta_{B.7}(s, \gamma, O(\delta'')), \delta_{B.8}(s, \gamma, O(\delta'')))$, $\tau = \tau'/2$, and $k = k_{B.8}(s, \gamma, O(\delta''))$. Then with probability at least $1 - \tau'/2$,

$$\mathcal{M}(F_G, \tilde{S}_G) \leq \mathcal{M}(F_G, S_G) + \mathcal{M}(S_G, \tilde{S}_G) \leq O(\delta''n) + O(\delta''n) = O(\delta''n).$$

The same inequality holds for the other graph H . Thus, with probability $1 - \tau'$,

$$\begin{aligned} &|\mathcal{M}(F_G, F_H) - \mathcal{M}(\tilde{S}_G, \tilde{S}_H)| \\ &\leq |\mathcal{M}(F_G, F_H) - \mathcal{M}(F_H, \tilde{S}_G)| + |\mathcal{M}(F_H, \tilde{S}_G) - \mathcal{M}(\tilde{S}_G, \tilde{S}_H)| \\ &\leq \mathcal{M}(F_G, \tilde{S}_G) + \mathcal{M}(F_H, \tilde{S}_H) \leq O(\delta''n) + O(\delta''n') = \delta''\tilde{n}. \end{aligned}$$

\square

B.3 Approximation algorithm

Finally, we show that the distance between two graphs can be well approximated by the minimum value of a matching between corresponding sketches. First, we need to show the following.

Lemma B.10. *Let G and H be s -rooted forests. Then, $d(G, H) \leq 2s \cdot \mathcal{M}(F_G, F_H)$.*

Proof. Let f^* be the optimal flow function achieving $\mathcal{M}(F_G, F_H)$. By Lemma B.5, we assume that every value of f^* is 0 or 1. Therefore, we regard the flow function as a matching: Let $F_G = (\mathbf{1}, \{\perp, \mathbf{u}_1, \dots, \mathbf{u}_{c(G)}\})$ and $F_H = (\mathbf{1}, \{\perp, \mathbf{v}_1, \dots, \mathbf{v}_{c(H)}\})$. Then, consider a bipartite graph such that the left part consists of $\{G^{(i)}\}$, the right part consists of $\{H^{(j)}\}$, and there is an edge between $G^{(i)}$ and $H^{(j)}$ iff $f^*(\mathbf{u}_i, \mathbf{v}_j) = 1$. Then, this graph forms a (partial) matching.

Using f^* , we construct a sequence of modifications to transform G to H . For each $G^{(i)}$ with $f(\mathbf{u}_i, \perp) = 1$, we remove all the edges in $G^{(i)}$. For each $H^{(j)}$ with $f(\perp, \mathbf{v}_j) = 1$, we remove all the edges in $H^{(j)}$.

Consider a pair $G^{(i)}$ and $H^{(j)}$ for which $f(\mathbf{u}_i, \mathbf{v}_j) = 1$. Let $\mathcal{T}_{G^{(i)}}$ and $\mathcal{T}_{H^{(j)}}$ be the set of subtrees in $G^{(i)}$ and $H^{(j)}$, respectively. From the definition, we can choose $\|\text{Freq}(G^{(i)}) - \text{Freq}(H^{(j)})\|_1$ sets of subtrees $\mathcal{T}'_{G^{(i)}} \subseteq \mathcal{T}_{G^{(i)}}$ and $\mathcal{T}'_{H^{(j)}} \subseteq \mathcal{T}_{H^{(j)}}$ in total so that $\mathcal{T}_{G^{(i)}} \setminus \mathcal{T}'_{G^{(i)}}$ and $\mathcal{T}_{H^{(j)}} \setminus \mathcal{T}'_{H^{(j)}}$ are isomorphic.

The total number of edge modifications is bounded by

$$\begin{aligned} & \sum_{\mathbf{u}_i: f(\mathbf{u}_i, \perp)=1} s \deg(\text{root}(G^{(i)})) + \sum_{\mathbf{v}_j: f(\perp, \mathbf{v}_j)=1} s \deg(\text{root}(H^{(j)})) \\ & + \sum_{(\mathbf{u}_i, \mathbf{v}_j): f(\mathbf{u}_i, \mathbf{v}_j)=1} s \|\text{Freq}(G^{(i)}) - \text{Freq}(H^{(j)})\|_1 \\ & \leq 2s \cdot \mathcal{M}(F_G, F_H). \end{aligned}$$

□

Now, we prove Lemma 4.1. We show that the following algorithm is a tester for forest-isomorphism.

Lemma B.11 (Restatement of Lemma 4.1). *There exists $\eta = \eta_{4.1}(s, \varepsilon', \gamma)$ such that for any input with $\frac{\tilde{n}}{n}, \frac{\tilde{n}}{n'} \in [1 - \eta, 1]$, the procedure **TestRootedForest** correctly decides $d(G, H) = 0$ or $d(G, H) \geq \varepsilon' \tilde{n}$ with probability at least $1 - \tau'$. Here η is a polynomial in ε', γ . Regarding that s is constant, the query complexity is a polynomial in $\gamma, \varepsilon', \tau', R$. Denote by $q_{\text{random}}^{4.1}(s, \gamma, \varepsilon', \tau')$ the number of random vertex queries the procedure invokes. Then $q_{\text{random}}^{4.1}$ is a polynomial in $\gamma, \varepsilon', \tau'$.*

Proof. Set $\eta = \eta_{B.2}(\delta)$. Combining Lemmas B.6 and B.10, the correctness of **TestRootedForest** can be proven as follows: If $d(G, H) = 0$, $\widetilde{M} \leq \mathcal{M}(F_G, F_H) + \delta'' \tilde{n} = \delta'' \tilde{n}$ (with probability $1 - \tau'$). On the other hand, if $d(G, H) \geq \varepsilon' \tilde{n}$, $\widetilde{M} \geq \mathcal{M}(F_G, F_H) - \delta'' \tilde{n} \geq (\varepsilon'/(2s) - \delta'') \tilde{n} > \delta'' \tilde{n}$.

The query complexity of **TestRootedForest** is polynomial in $\delta, \tau, q_{\text{freq}}, R, k^{t(s)}$. Since parameters δ, τ, k are polynomials in $\gamma, \delta'' = O(\varepsilon'/s), \tau'$, and q_{freq} is a

Algorithm 4 tests whether $d(G, H) = 0$ or $d(G, H) \geq \varepsilon' \tilde{n}$, with probability at least $1 - \tau'$, given $\tilde{n}, s, R, B, \gamma \geq 1$, $\varepsilon', \tau', \eta \in (0, 1)$ and R -good s -rooted forest with root degree in $(B, \gamma B]$ with $\frac{\tilde{n}}{n}, \frac{\tilde{n}'}{n'} \in [1 - \eta, 1]$ for $n = |V(G)|$ and $n' = |V(H)|$.

```

1: procedure TestRootedForest $_{\varepsilon', s, \tau', \eta, \gamma, R}(G, H, \tilde{n}, B)$ 
2:   Set  $\delta'' = O(\varepsilon'/s)$ .
3:   Choose parameters  $\delta, \tau, q_{\text{freq}}, k$  in Lemma B.6 according to parameters
    $\gamma, \delta'', \tau'/2, R$ .
4:   Compute  $\tilde{S}_G = \widetilde{\text{Sketch}}_{\delta, \tau, q_{\text{freq}}, R, k}(G)$  and  $\tilde{S}_H = \widetilde{\text{Sketch}}_{\delta, \tau, q_{\text{freq}}, R, k}(H)$ .
5:   Compute  $\tilde{M} = \mathcal{M}(\tilde{S}_G, \tilde{S}_H)$  by a min-cost flow algorithm.
6:   if  $\tilde{M} < \delta'' \tilde{n}$  then
7:     return YES
8:   else
9:     return NO

```

polynomial in $\gamma, \delta'', \tau', R$, the query complexity is a polynomial in $\gamma, \varepsilon', \tau', R$. We invoke random vertex queries $O(q_{\text{loop}})$ times for $q_{\text{loop}} = q_{\text{loop}}^{B.1}(k, s, \delta, \delta'_{B.2}(\delta), \tau)$, and therefore $q_{\text{random}}^{4.1}$ is a polynomial in $\gamma, \varepsilon', \tau'$. \square

C Missing Parts of Section 5

C.1 Missing proofs from Section 5

Proof (of Lemma 5.2). Note that $\frac{1+\mu}{1-\mu} < \gamma$. For a tree T in G , let p_T be the probability that T is on the (α, γ, μ) -boundary and $d'_T = \deg(\text{root}(T))$. Note that T is on the (α, γ, μ) -boundary if and only if $\alpha \in [\frac{\gamma^i}{d'_T(1+\mu)}, \frac{\gamma^i}{d'_T(1-\mu)}]$ for some i . Let $f_T(x) := |[1, \gamma] \cap [\frac{x}{d'_T(1+\mu)}, \frac{x}{d'_T(1-\mu)}]|$. Then, $p_T = \sum_{i \geq 1} f_T(\gamma^i) / (\gamma - 1)$ since the intervals $\{[\frac{\gamma^i}{d'_T(1+\mu)}, \frac{\gamma^i}{d'_T(1-\mu)}]\}_{i \geq 1}$ are disjoint as $\frac{1+\mu}{1-\mu} < \gamma$.

From the definition, if $\frac{x}{d'_T(1-\mu)} \leq 1$ or $\frac{x}{d'_T(1+\mu)} \geq \gamma$, then $f_T(x) = 0$ and otherwise $f_T(x) > 0$. Further, if $f_T(x) > 0$, then $f_T(\gamma^2 x) = 0$ since this implies $x > d'_T(1-\mu)$ and $\gamma^2 x > d'_T(1-\mu)\gamma \cdot \gamma > d'_T(1+\mu)\gamma$. It follows that $\#\{i \in \mathbb{N}_{<L+1} \mid f_T(\gamma^i) > 0\} \leq 2$. The value of $f_T(x)$ is maximized when $\frac{x}{d'_T(1-\mu)} = \gamma$. Thus, $f_T(x) \leq \gamma - \frac{1-\mu}{1+\mu}\gamma \leq 2\gamma\mu$, and we have $p_T \leq 4\gamma\mu$.

Therefore, we have $\mathbf{E}_\alpha[B_{\alpha, \gamma, \mu}(G)] = \sum_T p_T |V(T)| \leq 4\gamma\mu n$. By Markov's inequality, the lemma holds. \square

Proof (of Lemma 5.3). For a vertex $v \in V(G^{[0]}) \cup \dots \cup V(G^{[L]})$, let T be a tree with $v \in T$. If T contains no high-degree vertex, the procedure Which decides that $v \in G^{[0]}$. This output is correct.

Suppose that T contains a high-degree vertex u . Set $\delta = O(\mu/(\gamma^2 R))$ and $q = q_{5.3}(\gamma, \mu, R, \tau) = q_{3.3}(\delta, \tau)$. From Lemma 3.3 and the definition of an R -good tree, $|\deg_q(u) - \deg(u)| \leq \delta R \deg(u)$ holds with probability $1 - \tau$. If $v \in G^{[0]}$

or $v \in G^{[1]}$, the output is correct since $\delta R \cdot \alpha\gamma < 1/2$. Suppose that $v \in G^{[i]}$ for $i \in [2, L]$. Since $G^{[i]}$ is the union of trees that are not on the (α, γ, μ) -boundary, $(1 + \mu)\alpha\gamma^i < \deg(u) < (1 - \mu)\alpha\gamma^{i+1}$ holds. Thus, we obtain $\alpha\gamma^i < \widetilde{\deg}(u) < \alpha\gamma^{i+1}$. \square

Proof (of Lemma 5.4). The lemma follows from Lemmas 3.4 and 5.2. \square

Proof (of Lemma 5.6). Let A_k be the event that when running the procedure $\text{Random}_q(G, i)$, we pick up vertices of $(V(G^{[0]}) \cup \dots \cup V(G^{[L]})) \setminus V(G^{[i]})$ k times and pick up a vertex of $V(G^{[i]})$ and then, the procedure Which_q outputs always the correct value. Set $t = O(\log(1/\tau)/\delta)$. It is sufficient to show that $\Pr[A_0 \vee \dots \vee A_t] \geq 1 - \tau$ holds by appropriately choosing parameters. Let $\tau' = O(\delta\tau)$, $\lambda = \lambda_{5.6}(\delta, \tau) = O(\delta\tau)$, $q = q_{5.3}(\gamma, \mu, R, \tau')$. Then,

$$\begin{aligned} \Pr[A_0 \vee \dots \vee A_t] &= \sum_{k=0}^t ((1 - \lambda - \delta)(1 - \tau'))^k \cdot \delta(1 - \tau') \\ &\geq \delta(1 - \tau') \sum_{k=0}^t (1 - \lambda - \delta - \tau')^k \\ &\geq \delta(1 - \tau') \cdot (1 - \tau') / (\lambda + \delta + \tau') \geq 1 - \tau. \end{aligned}$$

\square

C.2 Tester for isomorphism of s -bounded-degree forests

We consider a forest-isomorphism tester for s -bounded-degree forests. As mentioned in Section 4, we need to make a tester for two forests containing different number of vertices. Using the result in [10], we can construct such a tester.

Lemma C.1. *There exists a procedure such that the following holds: For any $\varepsilon', \tau \in (0, 1)$ and $s \geq 1$, there exists $\eta = \eta_{C.1}(\varepsilon')$ such that for any s -bounded-degree forests G and H with $\frac{\tilde{n}}{n}, \frac{\tilde{n}'}{n'} \in [1 - \eta, 1]$, where $n = |V(G)|$ and $n' = |V(H)|$, the procedure correctly decides $d(G, H) = 0$ or $d(G, H) \geq \varepsilon' \tilde{n}$ with probability at least $1 - \tau$. The query complexity depends only on ε' and τ .*

Proof. We use the similar notion in [10]. For $s, k \geq 1$, let $N(s, k)$ be the number of rooted graphs whose degree is at most s and radius is at most k . Suppose that the $N(s, k)$ rooted graphs are numbered from 1. Let $N_i(s, k)$ be the i -th graph. In addition, for a vertex v in G , let $B_G(v, k)$ be the subgraph rooted at v that is induced by all vertices of G that are at distance at most k from v . Let $\text{Dist}_k(G)$ be the $N(s, k)$ -dimensional vector whose i -th element is the number of vertices v in G such that $B_G(v, k)$ is isomorphic to $N_i(s, k)$. Let $\text{Freq}_k(G) = \text{Dist}_k(G)/n$. The main result of [10] is as follows.

Theorem C.2 ([10]). *For any $\varepsilon' \in (0, 1)$ and $s \geq 1$, there exists $D = D_{C.1}(\varepsilon', s)$ and $\delta = \delta''_{C.1}(\varepsilon', s)$ such that if two forests G and H containing equal vertices are ε' -far from isomorphic, then $\|\text{Freq}_D(G) - \text{Freq}_D(H)\|_1 > \delta''$ holds.* \square

Without loss of generality, assume that $n \geq n'$. Let H' be a graph consisting of H and $(n - n')$ isolated vertices. We will prove the following claim.

Claim. If $\frac{\tilde{n}}{n}, \frac{\tilde{n}}{n'} \in [1 - \eta, 1]$, $\|\text{Freq}_D(H') - \text{Freq}_D(H)\|_1 \leq 4\eta$.

Using the claim, we can prove the lemma as follows. Set $\eta = O(\lambda)$. From the triangle inequality, $\|\text{Freq}_D(G) - \text{Freq}_D(H)\|_1 \geq \|\text{Freq}_D(G) - \text{Freq}_D(H')\|_1 - \|\text{Freq}_D(H') - \text{Freq}_D(H)\|_1$. If G and H are ε' -far from isomorphic, then $\|\text{Freq}_D(G) - \text{Freq}_D(H)\|_1 \geq \lambda - O(\lambda) = \lambda/2$ holds by Theorem C.2 and the above claim. Since we can approximate Freq by randomly sampling vertices and performing the BFS, we can test whether $\text{Freq}_D(G) = \text{Freq}_D(H)$ or $\|\text{Freq}_D(G) - \text{Freq}_D(H)\|_1 \geq \lambda/2$ with high probability.

We prove the claim. By the condition, $|1 - n/n'| \leq 2\eta$. Suppose that an isolated vertex is indexed one in the $N(s, D)$ -dimensional vector. Let z be the number of isolated vertices in H . Then,

$$\begin{aligned} & \|\text{Freq}_D(H') - \text{Freq}_D(H)\|_1 \\ &= \|\text{Dist}_D(H')/n - \text{Dist}_D(H)/n'\|_1 \\ &= |(z + n - n')/n - z/n'| + \sum_{2 \leq i \leq N(s, k)} |\text{Dist}_D(H)[i]/n - \text{Dist}_D(H)[i]/n'| \\ &\leq |1 - n'/n| + z|1/n - z/n'| + (\|\text{Dist}_D(H)\|_1 - z)|1/n - 1/n'| \\ &\leq |1 - n'/n| + |1 - n'/n| \leq 4\eta. \end{aligned}$$

□

We denote the procedure in Lemma C.1 by $\text{TestBoundedDegreeForest}_{\varepsilon', s, \tau, \eta}(G, H)$.

C.3 Proof of Theorem 1.1

Now, we prove Theorem 1.1. We show a tester for forest-isomorphism for s -forests in Algorithm 5. Let $\text{TestForestOfSameType}_{\varepsilon', s, \tau, \eta, \gamma, R}(i, G, H, \tilde{n})$ be the algorithm that runs $\text{TestBoundedDegreeForest}_{\varepsilon', s, \tau, \eta}(G, H)$ if $i = 0$, and runs $\text{TestRootedForest}_{\varepsilon', s, \tau, \eta, \gamma, R}(G, H, \tilde{n}, \alpha\gamma^i)$ otherwise.

We use the procedure TestIsomorphism in Algorithm 5 as a tester. It is sufficient to show the following theorem.

Theorem C.3. *The procedure $\text{TestIsomorphism}_{\varepsilon, \tau}(G, H)$ outputs the correct value with probability $1 - \tau$ with query complexity $\text{polylog}(n)$.*

Proof. We first calculate the probability that the procedure TestIsomorphism returns the correct value. In Line 7, $|V(G_{s, \alpha, \gamma, \mu, R}^{[L+1]})| \leq \lambda n$ and $|V(H_{s, \alpha, \gamma, \mu, R}^{[L+1]})| \leq \lambda n$ hold with probability $1 - O(\tau)$ by Lemma 5.4 and this is assumed in the following. In Line 10, both $|\tilde{z}_{G, i} - |V(G^{[i]})|| \leq \delta n$ and $|\tilde{z}_{H, i} - |V(H^{[i]})|| \leq \delta n$ hold for all i with probability $1 - O(\tau)$ by Lemma 5.5 and the union bound. Therefore, if $\tilde{z}_{G, i} \geq 2\delta n$, then $|V(G^{[i]})| \geq \delta n$ holds. (The same thing holds for $H^{[i]}$ as well.) Thus, from Lemma 5.6, we can provide the random vertex query to

Algorithm 5 returns YES if $d(G, H) = 0$ and NO if $d(G, H) \geq \varepsilon n$ with high probability, given two s -forests G, H and $\varepsilon, \tau \in (0, 1)$.

```

1: procedure TestIsomorphism $_{\varepsilon, \tau}(G, H)$ 
2:   Let  $\gamma = 2s$ ,  $L = O(\log n / \log \gamma)$ .
3:   Let  $\varepsilon' = O(\varepsilon/L)$ ,  $\tau' = O(\tau/L)$ ,  $\eta = O(\min(\eta_{B.2}(s, \varepsilon', \gamma, \tau'), \eta_{C.1}(\varepsilon)))$ ,
4:    $\delta = O(\min(\eta, \varepsilon'))$ ,  $\tau'' = O(1/(Lq_{\text{random}}^{4.1}(s, \gamma, \varepsilon', \tau')))$ ,  $q = q_{5.6}(\delta, \tau'')$ ,
5:    $\lambda = \lambda_{5.6}(\delta, \tau'')$ ,  $R = O(s/\lambda)$ ,  $\mu = O(\lambda/\gamma)$ ,
6:    $\beta_0 = 1/2$ ,  $\beta_{L+1} = 2\lambda$ ,  $\beta_1, \dots, \beta_L = (1 - \beta_0 - \beta_{L+1})/L$ .
7:   Choose  $\alpha \in [1, \gamma]$  uniformly at random.
8:   for  $i = 0, \dots, L$  do
9:     Let  $q_{\text{loop}} = q_{\text{loop}5.5}(\delta, \tau')$  and  $q_{\text{which}} = q_{\text{which}5.5}(\delta, \tau')$ .
10:    Compute  $\tilde{z}_{G,i} = \text{Size}_{q_{\text{loop}}, q_{\text{which}}}(G, i)$  and  $\tilde{z}_{H,i} = \text{Size}_{q_{\text{loop}}, q_{\text{which}}}(H, i)$ .
11:    if  $|\tilde{z}_{G,i} - \tilde{z}_{H,i}| > 2\delta n$  then
12:      return NO
13:    if  $\tilde{z}_{G,i} + \tilde{z}_{H,i} < \varepsilon' n$  then
14:      continue
15:    Let  $\tilde{n} = \max(\tilde{z}_{G,i}, \tilde{z}_{H,i}) + \delta n$ .
16:    Invoke the procedure  $\text{TestForestOfSameType}_{(\beta_i \varepsilon), s, (\beta_i \tau), \eta, \gamma, R}(i, G^{[i]}, H^{[i]}, \tilde{n})$ 
      with providing the random vertex query by  $\text{Random}_q(\cdot, i)$ .
17:    if the returned value is NO then
18:      return NO
19:   return YES

```

$G^{[i]}$ and $H^{[i]}$ correctly in Line 16 for every possible i with probability $1 - O(\tau)$. Here, the procedure Random invokes the procedure Which $O(1/(\delta\tau''))$ times. Again, in what follows, we assume these happen.

Suppose that $d(G, H) = 0$. In this case, the procedure returns NO in Line 12 for some i with probability at most $O(\tau)$. Further, the procedure returns NO in Line 18 for some i with probability at most $O(\tau)$. Therefore, the procedure returns YES with probability $1 - O(\tau)$.

Next, suppose that $d(G, H) \geq \varepsilon n$. From Lemma 5.1, there exists $i \in \mathbb{N}_{\leq L+1}$ such that $d(G^{[i]}, H^{[i]}) \geq \beta_i \varepsilon n$ holds. However, since we assumed that $|V(G^{[L+1]})|, |V(H^{[L+1]})| \leq \lambda n$ and we set $\beta_i = 2\lambda$, $d(G^{[L+1]}, H^{[L+1]}) \geq \beta_i \varepsilon n$ will never hold. Thus, we can say that there exists $i \in \mathbb{N}_{\leq L}$ such that one of the following holds: (i) $||V(G^{[i]})| - |V(H^{[i]})|| > 4\delta n$, or (ii) $||V(G^{[i]})| - |V(H^{[i]})|| \leq 4\delta n$ and $d(G^{[i]}, H^{[i]}) \geq \beta_i \varepsilon n$. If (i) holds, the procedure returns NO in Line 12. If (ii) holds, $|V(G^{[i]})| + |V(H^{[i]})| \geq \beta_i \varepsilon n$ holds (otherwise there exists a sequence of modifications from $V(G^{[i]})$ into $V(H^{[i]})$), and thus, $\tilde{z}_{G,i} + \tilde{z}_{H,i} \geq \varepsilon' n$ with the appropriate choice of constant factor of the parameters. Thus, the procedure does not pass Line 12. By the assumption, the condition of Lemma B.11 (i.e., $\frac{\tilde{n}}{|V(G^{[i]})|}, \frac{\tilde{n}}{|V(H^{[i]})|} \in [1 - \eta, 1]$) is satisfied. The procedure $\text{TestForestOfSameType}$ returns NO in Line 18 with probability $1 - O(\tau)$.

Applying the union bound for all assumptions stated so far, the procedure $\text{TestForestOfSameType}$ outputs the correct value with probability $1 - \tau$.

Every parameter here is a polynomial in ε and $L = O(\log n/\varepsilon)$, whose exponent is up to $\text{poly}(t(s))$. Since $s = s_{3.1}(\varepsilon)$ is a constant, the query complexity in Line 10 is polynomial in $O(\log n)$. Consider the query complexity in Line 16. When $i = 0$, we invoke the procedure `TestBoundedDegreeForest` with constant parameters. Thus, the query complexity is constant. When $1 \leq i \leq L$, the query complexity of the procedure `TestRootedForest` is polynomial in the parameters. Therefore, the query complexity of `TestIsomorphism` is $\text{polylog}(n)$ in total. \square

D Proof of Theorem 1.3

In this section, we prove that every property is testable with query complexity $\text{polylog}(n)$.

Proof (of Theorem 1.3). Suppose that we are given an oracle access to the input graph G . Let \mathcal{F} be the family of graphs that satisfy a property P . Consider the following procedure:

1. Use the parameters defined in Line 2–6,9 of Algorithm 5 and choose $\alpha \in [1, \gamma]$ uniformly at random.
2. For each $i \in \mathbb{N}_{\leq L}$, compute $\tilde{z}_{G,i} = \text{Size}_{q_{\text{loop}}, q_{\text{which}}}(G, i)$. If $\tilde{z}_{G,i} \geq O(\varepsilon' n)$ for $1 \leq i \leq L$, compute $\widetilde{\text{Sketch}}(G_{s,\alpha,\gamma,\mu,R}^{[i]})$. Let $\tilde{S}_G^{[i]} = \text{ext}(\widetilde{\text{Sketch}}(G^{[i]}), M)$, where $\text{ext}(\cdot)$ is an extension of a weighted point set and M is a sufficiently large value. Similarly, if $\tilde{z}_{G,0} \geq O(\varepsilon' n)$, compute an approximation to $\text{Freq}_{D_{C.1}}(G^{[0]})$.
3. For each $H \in \mathcal{F}$ and $i \in \mathbb{N}_{\leq L}$, compute $z_{H,i} = |V(H^{[i]})|$ and $F_H^{[i]} = \text{ext}(\text{Freq}(H^{[i]}), M)$. Note that we know the full information of \mathcal{F} , and therefore, we do not need to make any query to $H \in \mathcal{F}$. Then, test isomorphism between G and H in the similar manner as in Line 11–18. If $|\tilde{z}_{G,i} - z_{H,i}| > 2\delta n$, then regard that G and H are far from isomorphic. Otherwise, if $|\tilde{z}_{G,i} + z_{H,i}| \geq \varepsilon' n$, we test isomorphism by the sketch of $G^{[i]}$ and the weighted point set of $H^{[i]}$ as follows: If $1 \leq i \leq L$, compute $\mathcal{M}(\tilde{S}_G^{[i]}, F_H^{[i]})$. If it is sufficiently large, then regard that G and H are far from isomorphic. We perform the same thing if $i = 0$.
4. If there exists $H \in \mathcal{F}$ such that, for every $i \in \mathbb{N}_{\leq L}$, we do not regard that G and H are far from isomorphic, then return YES. Otherwise, return NO.

By the almost same argument as the proof of Theorem C.3, the procedure returns YES with high probability if $G \in \mathcal{F}$. We show that the procedure returns NO with high probability if G is ε -far from the property P . Let $F_G^{[i]} = \text{Freq}(G^{[i]})$. From Lemma B.10 and Lemma 5.1, for every $H \in \mathcal{F}$, $\mathcal{M}(F_G^{[i]}, F_H^{[i]}) = \Omega(\beta_i \varepsilon n)$ holds for some $1 \leq i \leq L$ (or $\|\text{Freq}_{D_{C.1}}(G^{[0]}) - \text{Freq}_{D_{C.1}}(H^{[0]})\|$ is sufficiently large for $i = 0$). Assume that $\mathcal{M}(F_G^{[i]}, \tilde{S}_G^{[i]})$ is sufficiently small. This happens with high probability. Then, from the triangle inequality, $\mathcal{M}(\tilde{S}_G^{[i]}, F_H^{[i]})$ is at least $\Omega(\beta_i \varepsilon n)$. The same argument holds for $i = 0$. Thus, the procedure will return NO with high probability.

The query complexity of this procedure is $\text{polylog}(n)$. \square

E Lower Bounds

In this section, we give an $\Omega(\sqrt{\log n})$ lower bound for testing forest-isomorphism and prove Theorem 1.2.

We first mention one technical issue to show lower bounds. Since H -isomorphism is a property which is closed under relabeling of vertices, we can assume that a tester for H -isomorphism does not exploit labels of vertices (see [7] for details). Instead, we assume that a tester obtains vertices by sampling vertices uniformly at random and only asks degrees and neighbors of sampled vertices.

We introduce several definitions for probability distributions. For two distributions \mathcal{D}_1 and \mathcal{D}_2 over S , the *total variation distance* between \mathcal{D}_1 and \mathcal{D}_2 is defined as $d(\mathcal{D}_1, \mathcal{D}_2) = \frac{1}{2} \sum_{i \in S} |\mathcal{D}_1(i) - \mathcal{D}_2(i)|$. For a set of elements S , we define $\mathcal{U}(S)$ as the uniform distribution over S .

We use the following lower bound as our starting point.

Lemma E.1 (Folklore). *Suppose that a probability distribution \mathcal{D} over $[s]$ is given as an oracle. That is, upon a query, we can sample an element from the distribution \mathcal{D} . We need $\Omega(\sqrt{s})$ queries to distinguish the case that $\mathcal{D} = \mathcal{U}([s])$ from the case that $\mathcal{D} = \mathcal{U}(S)$ for some $S \subseteq [s]$ with $|S| = \frac{s}{2}$.*

Now, we give a way of constructing a graph from a uniform distribution. To this end, we introduce a gadget. For an integer $k \geq 2$, let T_k be the *star graph* of k vertices. That is, the vertex set of T_k consists of a vertex v , called the *center vertex*, and vertices u_1, \dots, u_{k-1} connecting to v . For two integers N and k such that N is a multiple of k , we define T_k^N as the (disconnected) graph consisting of $\frac{N}{k}$ copies of T_k .

In what follows, we fix an integer N to be a huge power of two and $s = \log_2 N$ be an integer. From a uniform distribution \mathcal{U} over $S \subseteq [s]$, we construct its associated graph $G_{\mathcal{D}}$ by adding a copy of $T_{2^i}^N$ to $G_{\mathcal{D}}$ for each $i \in S$. Note that the number of vertices in $G_{\mathcal{D}}$ is $|S|N$, and $T_{2^i}^N$ is well-defined since $2^i \leq 2^s = N$.

Lemma E.2. *Suppose that a graph G is given as an oracle in the adjacency list model. We need $\Omega(\sqrt{s})$ queries to distinguish the case that $G = G_{\mathcal{U}([s])}$ from the case that $G = G_{\mathcal{U}(S)}$ for some $S \subseteq [s]$ with $|S| = \frac{s}{2}$.*

Proof. Given an oracle access to a probability distribution \mathcal{D} , which is guaranteed to be a uniform distribution over some set, we construct an oracle access to the graph G as follows.

Random-vertex query: We sample an element from \mathcal{D} and let i be the output.

Then, we construct a graph $T_{2^i}^N$ and return a random vertex in it. When we sample the same element i again, we reuse the same $T_{2^i}^N$.

Degree query: Let v be the specified vertex. Since v is a vertex returned by a random-vertex query, we know which $T_{2^i}^N$ contains the vertex v and how we have chosen v in $T_{2^i}^N$. Thus, we can return its degree.

Neighbor query: Let v and i be the specified vertex and index, respectively.

From the same reason as the previous case, we can return the i -th neighbor of v .

Note that the graph G behind the oracle we have designed is equal to $G_{\mathcal{U}(S)}$ when $\mathcal{D} = \mathcal{U}(S)$. Thus, from Lemma E.1, we have a lower bound of $\Omega(\sqrt{s})$ on the query complexity. \square

To obtain a lower bound for testing forest-isomorphism, we need to show that distinguishing two forests of the same number of vertices is hard. To address this issue, we use the following auxiliary lemma. For a graph G , we define $G^{\otimes 2}$ as the graph consisting of two copies of G .

Lemma E.3. *Suppose that a graph G is given as an oracle in the adjacency list model. For a subset $S \subseteq [s]$ with $|S| = \frac{s}{2}$, we need $\Omega(\sqrt{s})$ queries to distinguish the case that $G = G_{\mathcal{U}(S)}$ from the case that $G = G_{\mathcal{U}(S)}^{\otimes 2}$.*

Proof. The query-answer history of an algorithm is the subgraph obtained through the interaction to the oracle. As long as (the distribution of) the query-answer history is the same, (the distribution of) the output by the algorithm is the same (See, e.g., [7]). We can assume that the query-answer history does not have labels on vertices since we are assuming that algorithms do not depend on labels of vertices.

For each $i \in S$, $G_{\mathcal{U}(S)}^{\otimes 2}$ contains two copies of $T_{2^i}^N$. It is easy to see that the distribution of the query-answer history is the same as long as an algorithm does not hit vertices from both copies of $T_{2^i}^N$. Suppose that we have obtained a vertex from a copy of $T_{2^i}^N$ for some $i \in S$. The only way to obtain a vertex from the other copy of $T_{2^i}^N$ is querying random vertices. Thus from the birthday paradox, we need $\Omega(\sqrt{s})$ queries to obtain vertices from both copies of $T_{2^i}^N$ for some $i \in S$. \square

Since the number of vertices in $G_{\mathcal{U}([s])}$ and $G_{\mathcal{U}(S)}^{\otimes 2}$ is $sN = s2^s$, the value s is bounded from below by $\log n - \log s = \Omega(\log n)$, where $n = sN$. Thus, we have the following.

Corollary E.4. *Suppose that a graph G is given as an oracle in the adjacency list model. We need $\Omega(\sqrt{\log n})$ queries to distinguish the case that $G = G_{\mathcal{U}([s])}$ from the case that $G = G_{\mathcal{U}(S)}^{\otimes 2}$ for some $S \subseteq [s]$ with $|S| = \frac{s}{2}$.*

Now we show a lower bound for forest isomorphism. From Corollary E.4, we know that we need $\Omega(\sqrt{\log n})$ queries to distinguish the case $G = G_{\mathcal{U}([s])}$ from the case that $G = G_{\mathcal{U}(S)}^{\otimes 2}$ for some $S \subseteq [s]$ with $|S| = \frac{s}{2}$. In the former case, G is isomorphic to H . We finish the proof of Theorem 1.2 by showing that G and H are indeed far in the latter case.

The following lemma is useful to bound the distance between two graphs.

Lemma E.5. Let $G = (V_1, E_1)$ and $H = (V_2, E_2)$ be two graphs of n vertices. Then,

$$d(G, H) \geq \min_{\phi: V_1 \rightarrow V_2} \frac{1}{2} \sum_{u \in V_1} |\deg(u) - \deg(\phi(u))|,$$

where ϕ is over a bijection from V_1 to V_2 .

Proof. Let ϕ^* be a minimizer. For a vertex $u \in V_1$, we define $F(u)$ as the set of edges $(u, v) \in E_1$ incident to u such that $(\phi^*(u), \phi^*(v))$ is not an edge of E_2 . Clearly, $|F(u)| \geq |\deg(u) - \deg(\phi(u))|$ holds for every u . The lemma holds as $d(G, H) = \frac{1}{2} \sum_{u \in V_1} |F(u)|$. \square

Lemma E.6 (Lemma 3 of [13]). Let G_1 and G_2 be graphs. If some connected component C_1 in G_1 is isomorphic to a connected component C_2 in G_2 , then we can assume that C_1 is mapped to C_2 in an optimal bijection between G_1 and G_2 .

Lemma E.7. Let S be a subset of $[s]$ with $|S| = \frac{s}{2}$. Then $d(G_{\mathcal{U}([s])}, G_{\mathcal{U}(S)}^{\otimes 2}) \geq \frac{n}{8}$.

Proof. For notational simplicity, let $G = G_{\mathcal{U}([s])}$ and $H = G_{\mathcal{U}(S)}^{\otimes 2}$. From Lemma E.6, in the optimal bijection from G and H , we can assume that for each $i \in S$, $T_{2^i}^n$ in $G_{\mathcal{U}([s])}$ is mapped to the first copy of $T_{2^i}^n$ in $G_{\mathcal{U}(S)}^{\otimes 2}$. Let G' and H' be the graph obtained from G and H by removing these mapped vertices, respectively.

Now we consider the distance from G' to H' . We consider the loss caused by center vertices in stars of G' . Let u be a center vertex of a star in $T_{2^i}^n$ for some i . Then, u should be mapped to a vertex v in H' such that $\deg_{G'}(u) \leq \frac{1}{2} \deg_{H'}(v)$ or $\deg_{G'}(u) \geq 2 \deg_{H'}(v)$. From Lemma E.5, we have

$$d(G, H) \geq \frac{1}{2} \sum_{u: \text{center vertex in } G'} \frac{1}{2} \deg_{G'}(u) \geq \frac{n}{8}.$$

We have used the fact that the sum of degrees of center vertices of G' is $\frac{n}{2}$. \square

From the previous argument and Lemma E.7, we establish Theorem 1.2.